Networks and the evolution of complex phenotypes in mammalian systems.

Monzón Sandoval, Jimena

Award date:
2016

Awarding institution:
University of Bath

Networks and the evolution of complex phenotypes in mammalian systems.

Jimena Monzón Sandoval

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

October 2015

Table of Contents

Acknowledgments

First, I would like to particularly thank both Dr Araxi Urrutia and Dr Humberto Gutiérrez for the opportunity they give me to pursue a doctoral degree under their supervision, many thanks for the patience, guidance support and knowledge given to me during these years.

All my gratitude to Atahualpa Castillo Morales, friend, partner, colleague, confident and precious life companion who has shared with me, until now, a third of a life.

To all family members, many thanks, especially to my parents María Guadalupe Patricia Sandoval López and Miguel Salvador Monzón Bonilla, who always have trust me and encourage me to be who I am, and to whom I am most grateful.

I would also like express my endless gratefulness to my dear friends who have always inspire me and remained close to me however the distance, many thanks Mariana Reyes, Alejandra Zayas, Carlos Vargas, Rodrigo García and Leonardo Collado.

I am also very grateful to all past and present lab comrades, Jaime Tovar, Steve Bush, Nina Ockendon, Wei Wang, Atahualpa Castillo, Alin Acuña, Katy Mahler, Abdul Al-Zadjali and Bing Xia, together with dear friends at the Biology and Biochemistry department Dafina Angelova, Jennifer McDowald, Sergio Ancona, Cristina Carmona and everyone with whom I have shared ideas, jokes, complains, short trips, cake, sweets and drinks every now and then. My life in Bath would not have been the same without all of you.

Finally, the work presented here would have not been possible without the economic support of the Mexican National Council of Science and Technology (CONACyT), I acknowledge the funding that allow me to take part in this irreplaceable experience of studying a degree outside Mexico.

Gracias (Thanks).

Contributions

All the results and analysis presented in this thesis has been carried out by myself under the advice and support of my supervisors Dr Urrutia and Dr Gutiérrez except for the following contributions:

Chapter 1. Atahualpa Castillo organized GO enrichment analysis presentation results.

Chapter 2. Cell culture and microarray profiling experiments were performed by Sean Crampton, Laura McKelvey and Aoife Nolan under the supervision of Gerard O'Keeffe. Atahualpa Castillo normalized microarray expression of the aforementioned experiments.

Chapter 3. Human spheroid cultures and *yap1* knock down experiments were performed by Ruqyaiya Al Jabri, Tatsuo Miyamoto, Takeshi Senga and Makoto Furutani-Seiki. Atahualpa Castillo contributed to the analysis of RNA-seq data for *yap1* KD.

Chapter 4. Partial correlations and PIC analysis were performed by Atahualpa Castillo.

Abstract

Through a combination of multivariate and gene coexpression analysis of human brain transcriptome we provide evidence of an acute modular reorganization of the regulatory architecture of the brain transcriptome occurring at birth, reflecting the reassembly of new functional associations required for the normal transition from prenatal to postnatal development. Next, we showed a highly significant correlated expression and transcriptional clustering among immune-related genes in the developing and adult brain, which is not a common feature in other non-neural tissues, along with experimental evidence in which dissociated neurons were stimulated with TNF-alpha resulting in differentially expressed genes overrepresented in immune-related function, strongly suggesting a wide and coherent recruitment of entire immune related regulatory gene clusters by the nervous system. Then, we showed a marked difference in the gene sets coexpressed to the transcriptional co-activator *yap1* in the human brain between prenatal and postnatal developmental stages, each group of coexpressed genes displayed a significant enrichment of a series of quite different sets of functions, cellular localization and regulators, in agreement with the expected proliferative differences contrasting both developmental periods. We were able to identify changes in the close functional associates of *yap1* during this important developmental transition and potentially unveil novel roles in other functions.

Finally, through a comparative genomic analysis in mammalian species, we show parallel changes in gene family size and maximum lifespan that are not secondary to known correlates of lifespan, nor completely explained by phylogenetic effects. These changes preferentially occur in gene families associated to immune and defence response and remarkably to previously known gene variants associated to longevity in human populations, suggesting that underlying genetic adaptations of longevity and defence response mechanisms were in part brought about by changes in the number of gene copies within selected gene families.

List of abbreviations

A1C	primary auditory cortex
AMY	amygdaloid complex
CBC	cerebellar cortex
DFC	dorsolateral prefrontal cortex
HIP	hippocampus
IPC	posteroventral (inferior) parietal cortex
ITC	inferolateral temporal cortex
M1C	primary motor cortex
MFC	medial prefrontal cortex
OFC	orbital frontal cortex
S1C	primary somatosensory cortex
STC	posterior superior temporal cortex
STR	striatum
V1C	primary visual cortex
VFC	ventrolateral prefrontal cortex
pcw	post conceptional week
IS	Immune System
NS	Nervous System
DNA	deoxyribonucleic acid
RNA	ribonucleic acid
TCA	tricarboxylic acid
miRNA	micro RNA
shRNA	small hairpin RNA
TF	transcription factor
TFBS	transcription factor binding site
GO	Gene Ontology
KEGG	Kyoto Encyclopaedia of Genes and Genomes
TNF-alpha	tumor necrosis factor alpha
YAP1	yes-associated protein 1
TEAD	transcriptional enhancer activator domain containing protein
KD	knock down
HEK293	human embryonic kidney 293 cell line
RNA-seq	RNA sequencing
r	Pearson correlation coefficient

R	language and environment for statistical computing and graphics
WGCNA	Weighted Gene Coexpression Network Analysis
TO	topological overlap
GFS	gene family size
MLSP	maximum lifespan
Ei	encephalization index
Nr	neocortex to brain ratio
PIC	phylogenetically independent contrasts

Introduction

Functional correlates of transcriptome signatures

Transcriptome profiling can be analogous to taking a snapshot of the genome's dynamical behaviour at certain time and place. One of the aims of transcriptome analysis is to identify functional variations in the transcriptome in order to better understand the phenotypic differences between tissues, species, cells, individuals, developmental phases and disease states.

From the evolutionary point of view, an interesting observation is that conservation of expression patterns across species suggests selective pressure to maintain the transcriptional profile of individual tissues. Thus, for instance Brawand and colleagues analysed six different tissues (brain, cerebellum, heart, kidney, liver and testis) across 9 mammalian species and found conserved gene expression between tissues across species (Brawand et al 2011). This finding is consistent with other studies (Chan et al 2009, Dowell 2011, Merkin et al 2012, Shen et al 2015) also showing that the expression profile of a particular tissue is more similar to the same tissue in other species than to other tissue of the same species. A particularly remarkable finding of these studies is that nervous tissues display the most highly conserved expression pattern across species as compared to the remaining tissues, being testis transcriptome the tissue with the lowest degree of conservation.

At the cellular level, a recent work based on the analysis of the transcriptomes of 466 neural cells, managed to differentiate cellular populations in the human brain, without relaying in specific markers (Darmanis et al 2015). Interestingly, they also found that neurons have the largest number of expressed genes, when compared to the rest of cell types analysed (microglia, astrocytes, endothelial cells, oligodendrocytes and its precursor cells), and even managed to further identify seven distinct subpopulations of neurons demonstrating the power of single cell transcriptomics to increase our knowledge of the human brain at the cellular level.

The human body is composed of hundreds of different cell types (Hedges et al 2004). Each of them contains virtually the same genome but exhibit a particular set of specialized functions and morphologies. These phenotypic variations between tissues can only be explained by the specific segments of the genome being transcribed, i.e. their unique transcriptome (Mele et al 2015). One of the main divisions of the transcriptional profiles Mele et al. analysed was between solid and fluid tissues. Interestingly this study found that more than half of the total transcriptional output for the blood correspond to only three coding genes while other tissues such as kidney, heart, brain, colon and adrenal gland over a quarter of the total transcriptional output is given by just a few mitochondrion genes. In addition, when they compared the variation of different tissues across individuals of the same species (*Homo sapiens*), as expected from the conserved pattern of tissue specific patterns of expression between species (Brawand et al 2011, Chan et al 2009, Dowell 2011, Merkin et al 2012, Shen et al 2015), they found that most of the variation in the transcriptome is given by the diversity of tissues more than the differences between individuals. However, among the genes that exhibit most of the variation between individuals they found some disease candidate genes associated to the age, sex and ethnicity.

Together, these studies illustrate the diversity of transcriptional profiles at different levels (cell, cell subpopulation, cell population, tissues, individuals and species) and functional relevance at each of these levels in both normal and disease states.

Current transcriptional profiling approaches

The most widely used technologies to estimate the transcriptome are microarrays and RNA sequencing (RNA-seq).

Microarrays are tools that allows us to determine which genes are being transcribed at a certain point. Usually consist of a rectangular glass, silicon or plastic surface no larger than a few centimetres long where tiny ordered wells contain unique ~25 nucleotide sequences or probes bound to its surface, each capable of targeting by hybridization an individual transcript. Purified mRNA derived from a given biological sample can then be converted to complementary DNA (cDNA) and labelled with appropriate fluorophores. After hybridization with the microarray, it can then be

scanned and the relative expression of each gene can be estimated as the fluorescent intensity of the bound cDNA.

For RNA-seq on the other hand, fragmented purified mRNA is converted to cDNAs and sequenced in parallel. Typically, the order of the bases in determined by changes in light. The obtained reads are aligned to a reference genome to obtain the relative expression of a gene given as the number of reads that aligned to a particular region in the chromosomes (i.e. start to end of a known gene).

Microarrays were widely used during the first decade of this century, but RNA-seq technology has become more favoured as it provides a more accurate measurement of gene expression. One limitation of microarrays derives from the fact that transcript levels can only be measured from known genomic sequence information, sequences that are not printed in the microarray cannot be detected. In this sense RNA-seq permits to identify new transcripts, in addition, reads can be realigned once an updated (more complete or accurate) version of the genome has been generated. Moreover, RNA-seq can detect both coding and non-coding mRNAs. Another advantage of RNA-seq over microarray analysis is that reads can be filtered to keep only those that matched a unique location in the genome while in microarray analysis this is more difficult due to cross hybridization events. A common bias generated in RNA-seq data comes from the fact that a larger number of reads will map longer genes. However current normalization methods effectively correct for such biases. Additionally, alternative splicing events can be detected using RNA-seq if a deep enough sequencing has been performed. At the moment, there is a massive and growing collection of publicly available gene expression datasets generated by both microarray and RNA-seq platforms.

Differential gene expression and functional relevance in the nervous system

In addition to obtaining a complete gene expression profile, one of the main objectives of microarray analysis and RNA-seq technology is to identify genes that are expressed differently between biologically relevant samples. Because this approach typically

involves the screening of large numbers of genes, multiple testing corrections are needed to properly assess statistical significance in the observed changes.

In studies of the nervous system, most expression analyses have been done in the context of disease, comparing control versus affected individuals with the aim of obtaining prognostic or diagnostic gene signatures that could facilitate subsequent classification into risk groups. Differential expression analysis has been also helpful in understanding temporal and spatial differences between sexes, as well as between developmental periods or structures in the human brain (Berchtold et al 2008, Johnson et al 2009, Kang et al 2011, Lambert et al 2011, Oldham et al 2008).

For example, studies based on the transcriptional profiling of hypothalamus, hippocampus and frontal cortex in the developing rat showed that more than 95% of the expressed genes showed significant differences in their expression levels across development while more than 85% show significant differences in expression between regions (Stead et al 2006). Interestingly most of the spatial differences between regions increase during postnatal brain development.

Studies in human fetal brain showed that there are hundreds of genes differentially expressed between any two brain regions while there is little difference between the same regions in different hemispheres (Lambert et al 2011). During fetal development approximately 76% of genes are expressed and around 33% were differentially expressed between cortical areas (Johnson et al 2009) while around 86% of the genes expressed and 90% of them being differentially expressed either between regions or developmental ages (Kang et al 2011). Interestingly, here, most of the detected differences occur before birth. Together these studies suggest that most dynamic changes in gene expression in the brain occur during normal development.

Detection of differentially expressed genes

There are several methods to assess differential expression in transcriptome-wide studies, among them a simple t-test or its non-parametric version (Wilcoxon test) can be used in previously normalised data. For microarray data more elaborated methods including linear models of expression or rank products can be also used to detect differentially expressed genes (Breitling et al 2004, Diboun et al 2006, Ritchie et al

2015). For RNA-seq-based differential expression analysis methods are typically carried out using linear models of microarray analysis (limma), or packages implemented in R such as edgeR and DESeq (Anders & Huber 2010, Nikolayeva & Robinson 2014, Robinson et al 2010). There are many other methods/packages to assess differential gene expression form RNA-seq data, a fine comparison of the performance to detect differential expressed genes was done by Soneson and Delorenzi (Soneson & Delorenzi 2013).

Each of the different methods to detect differentially expressed genes will rely in specific assumptions of how the expression data is distributed, and this distribution will be influenced by the normalization methods used. It is worth noting on the other hand that, in order to detect differentially expressed genes, increasing the number of samples analysed is more important that increasing the sequencing depth, (especially when depths of over 20 million reads have been reached (Ching et al 2014)).

Gene coexpression and functional relevance

Gene coexpression refers to the coordinated expression of two or more genes across multiple samples. That is, if a pair of genes is (positively) coexpressed they tend to simultaneously increase and decrease their expression levels.

The main reason why gene coexpression analysis has been widely used is that it may reflect a functional association between genes, for example genes coding for components of the same protein complexes (Liu et al 2009), genes participating in closely related biological functions such as the mitochondrial respiratory function (Chen et al 2014), genes involved in the same metabolic pathway (Williams & Bowles 2004) or genes jointly regulated by the same transcription factors (Allocco et al 2004).

Previous studies done in *saccharomyces cerevisiae* demonstrated that any two genes displaying a high correlation in expression levels tend to share functional gene ontology annotations and transcription factors binding sites more often than expected by chance (Allocco et al 2004). A later study in *drosophila melanogaster* identified that only during early development genes with high coexpression patterns tend to share transcription factor binding sites more often than random gene pairs, and conversely

that genes with shared transcription factor binding sites showed higher coexpression than random gene pairs (Marco et al 2009).

Coexpression analysis can also be used in comparative studies. Thus, in a study done by Oldham et al. comparing the coexpression networks of the human and chimpanzee brain, they identified highly conserved gene coexpression networks between these closely related species, and even managed to identify conserved modules of coexpressed genes associated to specific regions such as cerebellum, cortex and caudate nucleus. Interestingly coexpression in the cortex was less conserved than the cerebellum, suggesting specific evolutionary patterns in the coexpression structure of the nervous system (Oldham et al 2006).

Even in plants such as the grapevine, previously unidentified genes that respond to heat were detected through a gene coexpression network analysis along with a cluster of coexpressed genes including several heat shock proteins (Liang et al 2014).

Taken together, these studies reveal a widespread functional signature that can be detected using gene coexpression networks, which has been proven useful in identifying gene groups or clusters of correlated genes that may participate in the same biological processes, molecular complexes, signalling or metabolic pathways.

In interpreting gene coexpression studies we have to bear in mind that gene coexpression may be a result of one or more reasons such as gene co-localization within chromosomes, epigenetic regulation, shared transcription factor regulation, miRNA regulation, cell type variation or even uncorrected batch effects (Gaiteri et al 2014).

Quantifying gene coexpression

There are several ways to measure gene coexpression, one of the most commonly used is through a simple Pearson correlation of the gene expression levels across a number of suitably chosen samples.

Spearman correlation can be also used to measure coexpression, but instead of measuring a linear relation between mRNA levels, it measures the relation between

the expression levels of a gene pair as a monotonic function of each other. Just like Pearson coefficients, Spearman coefficients can take values in the range from minus one to one.

Mutual information has been also used to measure coexpression. The mutual information refers to how much we can tell about a variable (the expression levels of a gene) given another variable (the expression levels of another gene). As this approach has been shown to yield virtually identical results to those obtained using Pearson or Spearman correlations of both correlation and mutual information analysis (Song et al 2012), and given that the permutation tests associated to the measure of mutual information are computationally costly, most studies of coexpression have been done based on correlations.

Other methods to measure coexpression include the use of Euclidean distances between the expression levels of a gene pair, partial correlations (de la Fuente et al 2004) as well as the use of jack-knife correlations, being the latter is a more robust estimate of coexpression and less sensitive to outliers than a normal correlation but is not frequently used due to its computational cost (Heyer et al 1999).

Gene coexpression network construction

For any given number of genes we can obtain a squared symmetric matrix containing correlation coefficients between every possible gene pair. From this matrix we can construct either a discrete gene coexpression network once we defined a certain threshold or keep the matrix as a set of continuous values to construct a weighted gene coexpression network.

Gene coexpression networks are often represented as graphs where genes are denoted by nodes and coexpression links between them by edges. As graphs, there are several characteristic and topological properties that define the graph and its elements, among them the most common include density, degree, clustering, centrality, shortest path between nodes and betweenness.

Biological properties of gene coexpression networks

One of the main properties of biological networks, coexpression networks included, is their singular topology described as scale free, which refers to the degree distribution of a networks following a power law, in other words, there are just a few highly connected nodes in the network, while the majority of the nodes have low connectivity (Barabasi & Oltvai 2004). Scale free networks are more resilient to perturbations except when a targeted removal of highly connected nodes is performed.

Within a network, highly connected or very central nodes are often called hubs. In *saccharomyces cerevisiae*, there seems to be a close connection between centrality and lethality in the protein-protein interaction network (Jeong et al 2001, Ning et al 2010), that is, hub proteins are three times more likely to be essential than less connected proteins. In coexpression networks, this centrality-lethality relationship seems to hold true, at least in yeast (Carlson et al 2006). Curiously, differentially expressed genes in depression disorders tend to be located in the periphery of gene coexpression networks (Gaiteri & Sibille 2011), which could reflect the impairment but not lethality associated to such diseases. As hinted by the work of Chu et al. differential connectivity pointed out genes relevant to a disease, in this case they compared cancer and healthy individuals that would otherwise be missed by differential expression analysis alone (Chu et al 2011).

Gene function inference can be improved using the shortest path within a biological network, for example if two non-adjacent genes in a network share a biological function, the genes located in shortest path are likely to have the same function (Zhou et al 2002).

Another interesting property of biological network it is their topological overlap (TO), this measures the number of shared neighbours between a pair of genes (Yip & Horvath 2007) and has been central to identifying biologically relevant modules or clusters of genes in the context of weighted gene coexpression network analysis (Eisen et al 1998, Langfelder & Horvath 2008, Oldham et al 2006). Additionally, TO serves as a filter to counter the effects of spurious or missing coexpression links between genes.

Remarkably, the increased dynamic range and accuracy of deep sequencing as compared with microarray platforms permits a better estimation of network properties, such as density, connectivity, centralization and heterogeneity (Iancu et al 2012), another reason to prefer RNA-seq over microarray transcriptome measurements whenever possible.

From gene coexpression networks to regulatory networks

Unlike gene regulatory networks, gene coexpression networks cannot be directed in the sense that we measure an association instead of the cause-effect relationship. From this point of view, multiple alternative gene regulatory relationships can fit within the same gene coexpression network. Although coexpression networks fail to identify causal relationships they nevertheless can be effectively sued to narrow down probable underlying regulatory relations.

For a given coexpression network a set of probable regulatory network can potentially be inferred. In this regard, efforts to reconstruct a regulatory network based on either time series, knock down and knock out experiments had been done but are still computationally and technically costly (Madar et al 2010, Pinna et al 2013, Pinna et al 2010). Interestingly, the predictive performance of these methods decreases for targets under the control of many regulators but not for regulators controlling several genes.

Differential coexpression and functional importance

Differential coexpression refers to the change of a coordinated pattern of expression of a set of genes between two or more conditions. It is worth mentioning again that changes in coexpression patterns of gene groups can happen without changes in the overall expression levels of individual genes.

These changes could consist of a switch from a coordinated pattern of expression in a gene group to a complete absence of coordinated expression and vice versa, or a rewiring of an existing coexpression architecture to an alternative one involving the exact same genes.

Just as coexpression analysis reveals groups of genes participating in a common process/pathway, differential coexpression analysis reveals restricted conditions under which groups of genes engage in a common process as they display coordinated expression in one condition but not the other.

In the context of disease, differential coexpression is often seen as an indication of a disrupted normal "healthy" network structure (de la Fuente 2010). At the cellular and physiological level there are many reasons why changes in coexpression may occur. Some of these variations may reflect different molecular interactions in response to certain conditions (environmental and/or genetic), variation in the regulation of gene groups (transcription factors, microRNA, epigenetic changes), normal processes that occur through development or aging, or as mentioned before, changes driven by a disease.

Thus, for instance, by comparing the gene coexpression networks of differentially expressed genes between obese and lean individuals, a recent study identified a subset of genes highly coordinated only in the obese network, NEGR1 was a hub gene in this network which has been found highly expressed in the hypothalamus where it appears to modulate synapse number in neurons, supporting an expected link between obesity and behaviour (Walley et al 2012).

A similar study focused on leukaemia-derived expression data identified several genes of the proteasome-ubiquitin pathways as well as the intracellular protein transport highly coordinated in healthy individuals and uncorrelated at the bone marrow of leukemia patients (Kostka & Spang 2004). Along the same lines, energy metabolism genes have been found deregulated in cancer samples; while genes involved in mitosis and those coding for collagens were found to be more tightly coordinated in cancer when compared to normal tissues (Choi et al 2005), revealing biologically relevant differences in the regulatory structure regulation between healthy and disease states.

Comparing young and old mice (16 and 24 months old respectively), Southworth et al. found some gene groups or modules that decrease their association (coexpression) as animals aged (Southworth et al 2009), among these modules this study identified a group of targets of the transcription factor NFkB, further illustrating differential

regulation associated to specific transcription factors, and also discovered that these genes with decreasing coexpression tend to be co-localized in the chromosomes, suggesting a possible influence of the local chromatin structure on the coordinated regulation of these genes.

Another example of age dependent gene coexpression changes comes from a study done by Gillis and Pavlidis where they analysed gene expression of different human tissues across four life stages (Gillis & Pavlidis 2009). Among the genes differentially coexpressed, the majority of hormone activity genes changed their coordination between prenatal and postnatal stages.

In all, coexpression analyses reveal the complex dynamical nature of the transcriptome, and help identify particular gene groups displaying coordinated activity in their levels of expression reflecting their functional association and specific patterns of spatiotemporal regulation.

Functional importance of gene family size variations

Gene families are groups of two or more genes with sequence similarity that arose through gene duplication. Through evolutionary time these gene families can further gain and/or loss members after rounds of gene duplication and deletion events and, consequently, different species can vary in the number of members that a particular gene family has. In this sense, gene families are composed by orthologous and closely related paralogs genes.

After a gene duplication event, the resulting gene copies can be lost o retained in a given lineage. Gene duplicates are more likely to change than singly copy genes (Han et al 2009), but are more likely to get lost (Demuth et al 2006). Even if sub and neofunctionalization are less frequent than the occurrence of deleterious mutations events, larger gene families provide the opportunity to evolve new or distinct but similar functions providing a source of diversity subject to adaptive forces.

A classic example of increasing functional diversity is given by the *hox* gene family. This gene family resulting from two rounds of whole genome duplication, is known to

control morphology and body axis formation across metazoans and as such has contributed to explain some distinct vertebrate-specific innovations (Holland & Garcia-Fernandez 1996, Soshnikova et al 2013). In the case of the sense of smell differences in the ability to detect odorants between mouse and human could be explained by the large differences in the size of the olfactory receptor gene family, one of the largest mammalian gene families (Young & Trask 2002). In a recent study we investigated gene family size variations in line with encephalization across several mammalian species and found numerous gene families comprising enriched in genes involved in cell-cell signalling, chemotaxis and immune system (Castillo-Morales et al 2014). More recent duplication and contraction events among primate species reveal human specific copy number variations, many of which consisted of genes involved in neurodevelopmental processes (Fortna et al 2004). These studies suggest that gene family size variations can contribute to the observed phenotypic differences between species.

Structure of the thesis

This thesis is composed of four self-contained chapters which can be read in any order however I suggest to read at least chapter 1 before 3.

Introduction to Chapter 1.

Human brain development is the result of a complex series of dynamic and adaptative processes, requiring the expression of specific genes at precise times and places. A vast majority of the genes will display temporal and/or spatial changes in gene expression throughout brain development support particular all the cellular functions required during this complex process (Johnson et al 2009, Kang et al 2011, Lambert et al 2011, Lister et al 2013, Oldham et al 2008).

Although discreet changes in individual genes can give us valuable insights into specific cellular functions, complex phenotypes are rarely the result of single genes acting in an isolation and are instead the consequence of the coordinated action of numerous molecular and genetic components (Hartwell et al 1999). One way to ensure this level of coordination is at the transcriptional level, so that genes involved in similar

biological pathways or processes will tend to display a similar patterns of expression reflecting their functional association (Eisen et al 1998, Homouz & Kudlicki 2013).

This coordination can be detected by measuring the correlation between the expression patterns of every gene pair across biological samples (i.e. tissues). Clustering analysis methods can then be applied to the resulting correlation matrices to detect groups or modules of correlated genes that may act together in particular molecular complexes, pathways, processes, or participate in similar regulatory and signalling circuits (Obayashi & Kinoshita 2011, Oldham et al 2006, Oldham et al 2008, Saris et al 2009, Torkamani et al 2010, Usadel et al 2009, Zhang et al 2012). As stated before, the modular structure of these coexpression networks not only reveal underlying functional interactions between coexpressed genes but also provide us clues on the regulatory architecture of the transcriptome as correlated genes are likely to be under the control of common transcriptional regulators (Marco et al 2009, Yu et al 2003).

During human brain development, the transcriptome has been shown to be structured into defined coexpression networks displaying distinct signatures of temporal expression (Kang et al 2011, Oldham et al 2008). In normal circumstances, these networks are assumed to be static and behave as single expression units co-varying in their level of expression across development possibly reflecting variations in their relative engagement during the developmental programme (Kang et al 2011). This assumption, however, has never been formally tested, at least to our knowledge.

There are many examples of genes that participate in more than one function, even when these biological functions appear to be completely unrelated (Harris et al 2004). This suggests a wide variety of possibilities in the assembly or rewiring of a coexpression network. Whether the observed variations in gene expression during normal development respond to concomitant changes in the level of expression of existing, and otherwise stable, networks of co-regulated genes or the potential regulatory reassembly of new functional clusters has never been explored.

To explore this possibility, in the first chapter of this thesis, we analysed genome wide expression data derived from the developing human brain at 14 developmental time points from 12 weeks post-conception through to 13 years of age spanning a total of 8

different cortical brain regions to determine whether the brain coexpression network constructed with these data is a mostly stable network or, conversely, a highly dynamic network.

Introduction to Chapter 2.

Several molecules and signalling components originally described as part of the immune system (IS) have been studied in the context of neural specific functions. The variety of processes in which some immune system genes are involved in the nervous system is not new, IS components have been linked to neural developmental processes such as survival, differentiation, dendritic growth, arborisation and axonogenesis as well as more complex processes including cognitive abilities (Carriba et al 2015, Galenkamp et al 2015, Gavalda et al 2009, Gutierrez & Davies 2011, Gutierrez et al 2005, Gutierrez et al 2013, Nolan et al 2011, O'Keeffe et al 2008, Twohig et al 2011). Remarkably, among genes with the highest variance gene expression during childhood in the human brain, an early stage characterized by synaptic plasticity, immune system genes are overrepresented along with key components of the nervous system development such as BDNF (Sterner et al 2012). In a wider scale analysis, immune system genes have been found overrepresented among gene families which size is associated to the degree of encephalization in mammalian species (Castillo-Morales et al 2014). Together these studies denote the involvement of immune system genes in the nervous system executing neural specific tasks. However, whether there has been a massive co-option of IS genes participating in neural-specific functions or if these cases are just isolated examples of IS genes involved in NS functions has remained mainly unexplored.

Functionally related genes, participating in the same processes, pathways and under the same regulators tend to share their patterns of gene expression and the patterns of coexpression have been shown a degree of conservation between species (Obayashi & Kinoshita 2011, Oldham et al 2006, Oldham et al 2008, Stuart et al 2003)

During the second chapter of the thesis we examined the degree the functional association of IS genes in the NS using as a proxy the coordination in their expression patterns in the brain at several brain structures and throughout the human lifespan.

Whether IS genes are particularly associated or not to other NS genes, and how the degree of association between IS genes compares to that of non-neural tissues, are addressed during this chapter.

Introduction to Chapter 3.

Classically studied as a downstream effector of the Hippo pathway, yes associated protein 1 (YAP1) participates in a variety of functions including regulation of organ size, proliferation, apoptosis, cell growth and differentiation (Camargo et al 2007, Lian et al 2010). More recently, YAP1 has also been found implicated in controlling the regulation of tissue tension in vertebrates (Porazinski et al 2015).

Briefly, when the Hippo pathway is activated, YAP1 is phosphorylated by LAST2 and retained in the cytoplasm. Conversely, when the Hippo pathway is repressed YAP1 can translocate to the nucleus and bind to transcription factors to promote the expression of hundreds of targets (Low et al 2014, Zhao et al 2008) .

YAP1 has also been shown to cross talk to other signalling pathways such as Notch, Wnt, EGF and TGFβ (Attisano & Wrana 2013, Fan et al 2013, Li et al 2012, Wang et al 2014). The stimulation and activation of these pathways may rely on cellular cues such as cell-cell contact, cell polarity, cellular stress, mechanotransduction, adhesion cues and extracellular signals (Low et al 2014, Zhao et al 2012). A connection between YAP1 and cell density has also been demonstrated. In high cell density conditions YAP1 is held in the nucleus while in lower density cell populations YAP1 translocate to the nucleus and promote transcription by binding the transcriptional regulator TEAD (Zhao et al 2007).

YAP1 protein is composed by several protein binding domains (WW, SH3, PDZ, TAD), allowing it to bind several proteins. For example, AMOT and LATS1 can bind to the WW domain of YAP1 (Oka et al 2008, Zhao et al 2011), while TEAD transcription factors binds to its TAD domain. Moreover, over a hundred proteins directly interact with YAP1 in HEK293 cells (Hergovich 2012), and either directly or indirectly YAP1 interacts with a large number of transcription factors including E2F, P73, SMADs, TBX5, CTNNB, ERB4 and RUNX2 (Alarcon et al 2009, Ehmer et al 2014, Kapoor et al 2014, Komuro et al 2003, Levy et al 2007, Rosenbluh et al 2012,

Strano et al 2001, Zaidi et al 2004). These large number of interactions evidence the complex molecular network of YAP1, which is also appears context dependent.

Coexpression analysis allows us to measure the association between genes across biologically relevant samples. Pairs of highly coexpressed genes have a higher tendency to participate in the same biological function (Allocco et al 2004) being this the main reason why a guilty by association approach has been widely use to annotate uncharacterized or poorly characterized genes (Wolfe et al 2005).

Furthermore, differential coexpression analysis allow us to identify species, age and disease specific changes in the coordinated patterns of gene expression (Amar et al 2013, Choi et al 2005, Chu et al 2011, Kostka & Spang 2004, Southworth et al 2009, Tesson et al 2010, Walley et al 2012). Differential coexpression analysis can identify transcriptome-wide changes that differential expression analysis alone cannot detect. These coexpression changes can reveal wider rearrangement of the underlying regulatory network in response to diverse stimuli or conditions (Tesson et al 2010).

During the 3rd chapter we aim to characterize the function of the transcriptional co-activator *yap1* at two developmental stages during brain development using a differential coexpression approach, in the context of brain developmental stages with diverse rates of cell proliferation. Given that cell proliferation is one of the most studied functions of *yap1* and that the differences in proliferation rates between prenatal and postnatal brain development, this presents an opportunity to identify changes in the close functional associates of *yap1* during this important developmental transition and potentially unveil novel roles of *yap1*.

Introduction to Chapter 4.

Lifespan has been defined as the time between we are born until the moment we die, and is limited in part by aging, characterized by a general decline of the organism with damage accumulation at the molecular, cellular and organ level, impairing the normal function and increasing the vulnerability to disease and death. It is clear however, that the resulting lifespan of an organism is a combination of both genetic and environmental factors (Kirkwood et al 2000, Walker 2011, Wensink et al 2012).

Comparative studies suggest that variations in maximum lifespan across species reflect intrinsic differences in the molecular machinery governing the ability of organisms to cope with age-related cellular damage and vulnerability to disease (Finch et al 2010, Harper et al 2007, Kirkwood et al 2000, Kourtis & Tavernarakis 2011, Perez et al 2009, Ricklefs 2010, Schumacher et al 2008). Ricklefs and Cadena, for example, detected and estimated a significant genetic contribution to lifespan on captive populations of wild animals (Ricklefs & Cadena 2008). Finch has reported that fibroblast form long-lived species are more resistant to several stressors and more likely to survive than those obtained from short-lived species (Finch et al 2010). Moreover it has been proposed that species specific cellular responses to stress and inflammation may contribute to the evolution of a longer life (Finch et al 2010, Iannitti & Palmieri 2011). Certain genes and pathways such as DNA damage response and ubiquitin pathways have been identified through species comparative studies relating accelerated protein evolution and mammalian longevity (Li & de Magalhaes 2013). Yet, the underlying genomic changes accounting for the differences in longevity across species remains largely unknown.

Although, through an evolutionary point of view gene number across species has remained largely constant the last 800 My, recent analyses have shown huge changes in GFS given by expansion and contractions events through time (Ashburner et al 2000, Demuth et al 2006, Fortna et al 2004, Hahn et al 2007a, Hahn et al 2007b, Hughes & Friedman 2004). As the fluctuations in GFS are particularly pronounced in gene families with specific functions, this suggest that changes in gene numbers within this gene families may reflect evolutionary responses to specific functional demands (Castillo-Morales et al 2014, Hahn et al 2007a, Hahn et al 2007b, Hughes & Friedman 2004, Kapheim et al 2015).

At the last chapter of this thesis we aim to gain insight in the genomic changes underlying the differences in longevity across mammalian species, to this end we use a comparative genomics strategy covering 28 sequenced mammals in order to identify associations between GFS variations and lifespan.

Chapter 1.

Modular reorganization of the global gene regulatory network architecture during perinatal human brain development.

Jimena Monzón-Sandoval[1, 2], Atahualpa Castillo-Morales[1, 2], Araxi O. Urrutia[2], Humberto Gutierrez[1]

[1] School of Life Sciences, University of Lincoln, Lincoln LN6 7DL, UK

[2] Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

JMS: jms52@bath.ac.uk

ACM: acm39@bath.ac.uk

AOU: a.urrutia@bath.ac.uk

HG: hgutierrez@lincoln.ac.uk

Abstract

During early development of the nervous system, gene expression patterns are known to vary widely depending on the specific developmental trajectories of different structures. Observable changes in gene expression profiles throughout development are determined by an underlying network of precise regulatory interactions between individual genes. Elucidating the organizing principles that shape this gene regulatory network is one of the central goals of developmental biology. Whether the developmental programme is the result of a dynamic driven by a fixed architecture of regulatory interactions, or alternatively, the result of waves of regulatory reorganization is not known. Here we contrast these two alternative models by examining existing expression data derived from the developing human brain in prenatal and postnatal stages. We reveal a sharp change in gene expression profiles at birth across brain areas. This sharp division between foetal and postnatal profiles is not the result of sudden changes in level of expression of existing gene networks. Instead we demonstrate that the perinatal transition is marked by the widespread regulatory rearrangement within and across existing gene clusters, leading to the emergence of new functional groups. This rearrangement is itself organized into discrete blocks of genes, each associated with a particular set of biological functions. Our results provide evidence of an acute modular reorganization of the regulatory architecture of the brain transcriptome occurring at birth, reflecting the reassembly of new functional associations required for the normal transition from prenatal to postnatal brain development.

Introduction

Development of the human nervous system is a complex and precisely regulated process that occurs over a prolonged period of time and depends on the precise temporal and regional coordination of complex patterns of gene expression.

As the developmental programme unfolds, groups of genes involved in a variety of functions change their level of expression in the brain at specific times in response to changing demands for a wide variety of specific cellular processes. During human nervous system development, up to 89.9% of expressed genes have been shown to be temporally differentially expressed between any two periods across regions, with 85.3% of genes being differentially expressed between any two periods across areas in the neocortex alone (Stead et al 2006, Sterner et al 2012).

Waves of intense variation in gene expression are particularly pronounced in specific stages of development. In the rat brain model, for instance, for most genes, the most dramatic changes in level of expression occur early in postnatal life (1–2 weeks) and plateau thereafter (Stead et al 2006).

However, complex phenotypes and indeed specific cellular functions are rarely the result of individual genes acting in isolation and are instead the result of a complex assembly of molecular and genetic components acting in concert (Hartwell et al 1999). As a result, genes involved in related biological pathways tend to display similar expression pattern reflecting their functional association (Eisen et al 1998, Homouz & Kudlicki 2013). This coordinated expression can be readily detected by looking at existing correlations in expression levels between groups of genes across a series of suitable chosen tissue samples. Clustering analysis based on coexpression patterns has been used to identify groups or modules of correlated genes that may form molecular complexes, pathways, or participate in common regulatory and signalling circuits (Obayashi & Kinoshita 2011, Oldham et al 2006, Oldham et al 2008, Saris et al 2009, Torkamani et al 2010, Usadel et al 2009, Zhang et al 2012). Apart from revealing functional interactions among groups of genes, gene coexpression also provides information on the regulatory architecture associated to a global expression profile as

co-expressed genes are likely to be under the concerted control of a common complement of transcriptional regulators (Marco et al 2009, Yu et al 2003).

During brain development, the transcriptome has been shown to be organized into distinct coexpression networks displaying clearly defined patterns of temporal expression (Kang et al 2011, Oldham et al 2008). In normal circumstances, these networks are assumed to behave as single expression units co-varying in their level of expression across development possibly reflecting variations in their relative engagement during the developmental programme (Kang et al 2011). This assumption, however, has never been formally tested, at least to our knowledge.

On the other hand, individual genes have the potential to participate in more than one separate and sometimes seemingly unrelated biological function (Harris et al 2004). Whether the observed variations in gene expression during normal development respond to concomitant changes in the level of expression of existing, and otherwise stable, networks of co-regulated genes or the potential regulatory reassembly of new functional clusters has never been explored.

In this study, we analysed genome wide expression data derived from the developing human brain at 14 developmental time points from 12 weeks post-conception through to 13 years of age spanning a total of 8 different cortical brain regions. Clustering analysis of expression profiles show that gene expression throughout development is divided into two clearly defined temporal domains before and after birth. This sharp division between foetal and postnatal profiles is not the result of acute changes in the level of expression of existing networks of co-regulated genes. Instead we demonstrate that the perinatal transition is marked by the widespread regulatory rearrangement within and across existing gene clusters, giving rise to the emergence of new functional groups. Our results demonstrate an acute regulatory reorganization of the brain transcriptome occurring specifically at birth and reflecting the reassembly of new functional associations potentially required during the transition from prenatal to postnatal brain development.

Methods

Expression data

RNA-seq RPKM normalized expression data summarized to genes was obtained from NIMH Transcriptional Atlas of Human Brain Development (Hawrylycz et al 2012, http://www.brainspan.org/). We selected 112 samples corresponding to 8 brain structures for which there was available data across 14 early stages. This resulted in eight cortical regions: Dorsolateral prefrontal cortex (DFC), Posteroinferior (ventral) parietal cortex (IPC), Anterior (rostral) cingulate (medial prefrontal) cortex (MFC), Orbital frontal cortex (OFC), Primary somatosensory cortex (area S1, areas 3,1,2) (S1C), Posterior (caudal) superior temporal cortex (area TAc) (STC), Primary visual cortex (striate cortex, area V1/17) (V1C) and Ventrolateral prefrontal cortex (VFC). Seven of the fourteen different developmental stages developmental stages correspond to post-conception weeks 12, 13, 16, 17, 21, 24 and 37. The other seven postnatal time points are 4 months after birth followed by 1, 2, 3, 8, 11 and 13 years of age. We selected only protein coding genes according to Ensembl version 77 and removed from the analysis all genes displaying zero variance across samples resulting in a total of genes (n = 18526).

Expression profile clustering analysis

To quantify similarity of expression profiles across brain structures and two developmental stages (prenatal and postnatal), we obtained the Pearson correlation coefficient (r) between the normalized average expression values per gene per structure of all possible pairs of expression profiles. We defined distance between any two expression profiles as 1- r and performed average linked hierarchical clustering analysis.

Coexpression network clustering analysis

To measure the degree of similarity in the coexpression structure of the same set of prenatal and postnatal brain regions, we compared the coexpression matrices (Pearson correlation matrices) between all possible pairs of genes across all prenatal and postnatal samples for each brain structure. More specifically, for each brain structure,

we obtained the coexpression matrix (defined as the Pearson correlation matrix between all genes) across all seven prenatal time points. We repeated the same procedure for all postnatal time points resulting in a total of 16 global coexpression matrices. We defined similarity between any two coexpression matrices as 1- r, where r represents the Pearson correlation coefficient between any two coexpression matrices. The resulting similarity indexes were used to perform a hierarchical clustering analysis.

Differential coexpression analysis

To quantify changes in the global pattern of coexpression in the perinatal boundary, we performed differential coexpression analysis as described by Tesson (Tesson et al 2010) based on a Weighted Gene Correlation Network Analysis (WGCNA) approach. Briefly; we calculated correlation coefficients for all possible gene pairs separately for the prenatal and postnatal period, obtaining one global correlation matrix for each stage. Then we computed the adjacency difference matrix using the soft threshold parameter $\beta = 6$ (in order to achieve the scale–free topology fitting index $R^2 > 0.8$). Next, hierarchical clustering was performed based on the Topological Overlap of the difference matrix. Finally, the dynamic tree cut function (implemented in R) was used in order to identify gene modules (minimum cluster size of 100 genes deep split = TRUE). Modules where merged when the module's eigengenes correlation was higher than r <= 0.9 (shown in colour bar of figure 3A).

Gene Ontology enrichment analysis

We downloaded gene ontology biological process (GO) annotations from Ensembl version 77 (http://www.ensembl.org/index.html), selected only those GO containing at least 150 genes with available expression data (n = 106) and performed an enrichment analysis for each of the modules detected in the differential coexpression analysis. Statistical significance was assessed by counting within each module the number of genes annotated to each GO term and compared these counts to those derived from 1000 random equally sized samples drawn from the whole gene population (Monte Carlo simulation) to derive a Z score with their corresponding p value, which was then adjusted for multiple testing using a Benjamin-Hochberg correction. The area proportional Venn and Euler diagram in figure 4B was created

using the *venneuler* function supported in R, where the area is proportional to the number of significantly enriched GO categories (FDR < 0.05) and a difference between observed and expected genes larger than one.

Programming and statistical software

All large scale calculations, numerical simulations and statistical analysis were carried out in R. The Differential coexpression analysis script was obtained from Tesson et al. (Tesson et al 2010) already implemented in R. Network visualization was implemented in Gephi using the inbuilt Fruchterman-Reingold layout.

Results

We started by asking whether the most prominent component of variance in gene expression during brain cortex development was due to regional or temporal variations in gene profiles. To this end we examined RNA-seq expression data obtained from the NIMH Transcriptional Atlas of Human Brain Development (Hawrylycz et al 2012, http://www.brainspan.org/). We selected 112 samples corresponding to 8 brain structures for which there was available data across 14 early stages (post-conception weeks 12, 13, 16, 17, 21, 24 and 37; 4 months after birth as well as 1, 2, 3, 8,11 and 13 years of age. After removing from the analysis all genes displaying zero variance across samples, resulting in a total of 18526 genes.

We next carried out a multiple component analysis splitting samples by either region or post conception age. Using the first and second components (together contributing to 68.83% of variance) we found no significant association between variations in gene expression and anatomical structure ($F_{8, 103} = 0.177$, $p = 0.994$, figure 1A). By contrast, the global expression pattern showed a highly significant association with developmental stage ($F_{1, 110} = 77.6$, $p = 2.069 \times 10^{-14}$, figure 1B) demonstrating a more prominent contribution of the developmental stage to the observed changes in gene expression than differences associated to different regions. Furthermore, when we split expression data into prenatal and postnatal samples, the association between expression profiles and these two developmental windows was even more pronounced ($F_{1, 110}$ $p = 9.998 \times 10^{-46}$, figure 1C).

These results show that the single greatest component of the variance corresponds to the developmental stage of the brain rather than anatomical structure. More specifically, these results reveal a distinctly pronounced transcriptional profile shift between prenatal and postnatal expression irrespective of brain region.

To directly test this possibility we assessed the transcriptional relatedness between all brain regions, averaging, for each brain region, prenatal and postnatal expression per gene, resulting in a total of 16 average expression profiles; one for each of the 8 brain regions at either prenatal or postnatal stages. Using these profiles, we calculated correlation matrices of pairwise comparisons followed by unsupervised hierarchical clustering. This analysis revealed two highly correlated expression profiles sharply dividing foetal and postnatal stages (figure 1D). These results show that any two brain regions are more similar to each other within each developmental window than they are to themselves across the perinatal boundary and demonstrate the existence of two distinct global expression patterns characterizing the prenatal and postnatal stage in nervous tissues irrespective of which brain structure they belong to.

As mentioned above, complex phenotypes are rarely the result of individual genes working in isolation and are, instead, the result of an assembly of molecular and genetic components acting in concert (Hartwell et al 1999). Consequently, genes involved in related biological pathways tend to display similar expression patterns reflecting their functional association (Eisen et al 1998, Homouz & Kudlicki 2013). Along these lines, the developmental brain transcriptome has also been shown to be organized into distinct coexpression networks each one displaying clearly defined patterns of temporal expression (Kang et al 2011, Oldham et al 2008).

The observed switch in the global expression profile sharply dividing the prenatal and postnatal developing human nervous system can be the result of two alternative underlying processes: A) a sudden change, during the perinatal boundary, in the overall level of expression of existing, otherwise cohesive, gene clusters (figure 2A) or B) An overall regulatory reassembly of new functional clusters (figure 2B).

One way to discriminate between these two possibilities is by looking at the similarity pattern in the coexpression structure of individual brain regions comparing both the prenatal and postnatal developmental stages (figure 2 A and B right).

To this end, we used a weighed gene coexpression network analysis approach (WGCNA) where the coexpression structure of the transcriptome of a given brain region can be represented as the Pearson correlation matrix of all possible pairs of genes across a number of developmental time points. Accordingly, we obtained the coexpression matrices of each brain region for prenatal and postnatal stages separately, resulting in a total of 16 different coexpression matrices: one for each of the 8 brain regions at either prenatal or postnatal stages. We conducted an average linkage hierarchical clustering analysis and defined similarity between any two coexpression matrices as (1 - r), where r is the Pearson coefficient derived from correlating any two coexpression matrices.

As shown in figure 2C, clustering of brain regional samples based on their coexpression structure shows a clear split between pre and postnatal samples, with coexpression structures within each developmental stage resembling more each other irrespective of brain region, than the same region resembling itself across these two developmental windows. This result demonstrates an overall reorganization of the coexpression structure of the brain transcriptome as the developmental program crosses the perinatal boundary (as shown in the schematic model of figure 2B), further revealing a sharp remodelling in the gene regulatory structure of the developmental programme between late prenatal and early postnatal human brain.

In order to quantify the pattern of regulatory changes occurring during the perinatal boundary, we conducted differential coexpression analysis as described by Tesson et al, (Tesson et al 2010). This method groups genes together when their correlations with the same sets of genes change between the different conditions. Briefly, we obtained the overall coexpression matrices for either prenatal or postnatal stages, each one comprising data from all 7 ages and 8 brain regions and obtained the difference matrix resulting from subtracting one from the other. A topological overlap matrix based on the differential coexpression matrix was then calculated followed by hierarchical clustering to identify modules of differentially coexpressed genes (figure

3A). This analysis identified 23 modules of differential coexpression ranging in size from 115 to 3021 genes (figure 3A and B). A close inspection of the correlation heatmaps of the resulting clusters confirms pronounced changes in the correlated structure of each module in the transition between prenatal to postnatal development with most modules displaying an overall increase in correlated activity in the postnatal stage (figure 3C). We quantified this effect by simply measuring the change in the average correlation of each module between pre and postnatal stages (figure 3D) and found that 17 out of 23 differentially coexpressed modules displayed a significant increase in correlated expression in the postnatal stage with 6 modules showing reduced correlated activity in the same developmental stage.

Together, these results demonstrate an overall reassembly of the regulatory structure of the brain transcriptome in the perinatal boundary (figure 2B), and that this reassembly is itself organized into discrete modules or clusters of genes undergoing intensive regulatory reorganization.

In order to test the functional significance of the observed modular reorganization of the brain transcriptome in the perinatal boundary, we asked whether each regulatory reorganization module targeted a specific set of biological functions. To this end we determined the number of gene ontology (GO) terms, within the biological process category, statistically overrepresented within each module and assessed the overlap in GO terms between modules. As shown in figure 4A, 19 out of 23 regulatory reorganization modules displayed significant enrichments in one or more specific biological processes. In the Venn-Euler plot of figure 4B, the area of each circle represents the number of enriched GO terms in each module and the overlap represents the relative proportion of overlapping GO terms between modules. As can be seen in the graph, each module targets an almost exclusive set of biological functions with rare functional overlap between modules. This result shows that the observed regulatory reorganization of the transcriptome in the perinatal boundary is organized into discrete regulatory remodelling networks targeting defined and almost non-overlapping sets of biological functions.

Within each module, the observed regulatory reorganization involved a combination of events of increased and decreased coordinated activity between individual gene

pairs. In the examples shown in figure 5, corresponding to modules M15 and M7, each module is represented as a graph, where nodes represent genes and edges represent an existing high correlation ($r > 0.95$) between the indicated pair of genes. Red edges represent prenatal-only high correlations, whereas blue edges represent postnatal-only high correlations. Module M15 shows an overall transition from high to low correlated activity for all involved gene pairs in the transition from prenatal to postnatal development. By contrast module M7 shows a transition from high to low coordination between genes in a subnetwork accompanied by a transition from low to high correlated activity in a second sub-network.

Taken together our results demonstrate an acute and modular regulatory reorganization of the brain transcriptome occurring at birth and reflecting the reassembly of new functional associations potentially required during the transition from prenatal to postnatal brain development.

Discussion

The development of the nervous system is a highly complex process, involving the coordinated regulation of thousands of genes. As the developmental program unfolds, gene expression patterns are expected to vary widely depending on the specific developmental trajectories of different tissues (Stead et al 2006, Sterner et al 2012).

Genes however do not act in isolation, as most cellular, physiological and developmental functions are the result of a complex assembly of molecular and genetic components acting in concert (Hartwell et al 1999). As a result, genes involved in related biological pathways tend to display correlated expression patterns reflecting their functional association (Eisen et al 1998, Homouz & Kudlicki 2013).

Developmental, regional or temporal variations in gene expression can therefore be understood as resulting from changes in the relative level of expression of existing, and otherwise cohesive, networks of co-regulated (functionally)-related genes. These changes can in principle take place as part of the normal dynamics of an otherwise fixed regulatory architecture (figure 2A). In other words, the regulatory architecture can be regarded as a constant feature of the developmental process and alterations of

this architecture could reflect regulatory instabilities resulting in abnormal development and pathological conditions.

In this regard, changes in the correlated status of groups or networks of genes have indeed been linked to regulatory dysfunctions associated to pathological conditions such as cancer and some neurodegenerations (Choi et al 2005, Miller et al 2008) or interpreted as an instance of genome instability associated to age-related functional decline (Southworth et al 2009).

On the other hand, individual genes have the potential to participate in more than one separate and sometimes seemingly unrelated biological functions (Harris et al 2004). Depending on the varying developmental requirements for specific sets of cellular functions, changing transcriptional profiles could alternatively reflect underlying changes in the regulatory architecture normally required at critical stages of development. These changes, in turn, are expected to result in an overall regulatory reassembly of new functional clusters of co-regulated genes.

In this study, we specifically asked whether, during the normal developmental process, changes in transcriptional profiles are the result of an otherwise stable regulatory architecture or, alternatively, the result of successive waves of regulatory reorganization taking place at critical stages of development. By analysing expression data of the human brain developmental transcriptome and comparing the relatedness of prenatal and postnatal expression profiles across several brain structures we found two distinct sets of transcriptional signatures sharply dividing the prenatal and postnatal window irrespective of brain region. While this result is consistent with both a constant and a developmentally variable underlying regulatory architecture, here we uncover evidence of intense regulatory reorganization occurring in the transition from prenatal to postnatal development. Using the coexpression matrices as a measure of the regulatory architecture and comparing across brain regions at either prenatal and postnatal stages, we found a sharp split in the regulatory architecture occurring at the perinatal boundary, so that, within each developmental window, any two brain regions are more similar to each other at the level of their coexpression structure than they are to themselves across these two developmental stages.

Using differential coexpression analysis, we further characterized the overall remodelling of the regulatory structure of the brain transcriptome at birth and found that this reassembly is itself structured into discrete modules or clusters of genes undergoing intensive regulatory reorganization.

In order to gain insights into the functional coherence of the observed modular reorganization of the brain transcriptome at birth, we asked whether these reorganization clusters targeted specific biological functions and if so, to which extent different clusters target different sets of biological functions. Gene ontology enrichment analysis revealed that each module targets a separate set of biological functions, with rare functional overlap between modules. This result shows that the observed regulatory reorganization of the transcriptome in the perinatal boundary is organized into discrete clusters targeting distinct sets of biological functions.

The transition from prenatal to postnatal development is marked by drastic changes in the physiological environment under which the developmental programme unfolds, not least the transition from intra to extra uterine conditions. Under these circumstances the organism faces the challenge of continuing with a normal developmental trajectory under a whole new set of environmental conditions. This transition entails the overall adaptation of the developmental programme to a wide range of new environmental variables.

This adaptation can conceivably demand the widespread remodelling of previously existing regulatory interactions within and between gene networks involved in a wide array of existing and/or emerging of cellular and developmental functions.

Here, we observed a shift in the overall coexpression structure of the developmental transcriptome of the brain revealing a corresponding widespread reorganization of the underlying gene regulatory circuitry. The fact that this transition is organized into discrete modules targeting distinct sets of biological functions strongly suggests the emergence of the new functional associations required for the normal transition from prenatal to postnatal brain development. Thus for instance, in the example shown in figure 5A, module M15 displays an overall reduction of correlated activity between these genes in the transition from prenatal to postnatal development. Interestingly, this

module is statistically enriched in genes involved in cell cycle, mitosis and cell proliferation functions. This reduction in the level of coordination between genes directly involved in proliferative functions could potentially reflect corresponding differences in the level of engagement of proliferative activity during prenatal and postnatal nervous system development. Indeed, cell proliferation is particularly pronounced during embryonic and late foetal stages of nervous system development as neuronal progenitor cells proliferate and their population expands to eventually differentiate into mature post-mitotic neurons. At postnatal stages, proliferative activity almost ceases for neural precursors and is virtually restricted to glial proliferation leading to a low, but sustained production of both astrocytes and oligodendrocytes throughout the adult brain (Guerout et al 2014, Rowitch & Kriegstein 2010).

Conversely module M7, shown in figure 5B is characterized by a transition from high to low coordination between genes in a subnetwork with a simultaneous transition from low to high correlated activity in a second sub-network. Not surprisingly this module is enriched in a large number of functions including gene expression, RNA splicing, translation, RNA and protein metabolic processes. Here is noteworthy the transitivity of the subnetworks in which gene reassemble and coordinate its expression patterns with other genes of the same module.

Conclusion

In all, we conclude that, during brain development, the pronounced changes in the genome wide expression profile observed in the perinatal boundary are the result or a regulatory reorganization of the developmental programme occurring at birth and reflecting the reassembly of new functional associations required for the normal transition from prenatal to postnatal nervous system development. We propose that these developmental changes in the patterns of coexpression reveal underlay demands of regulatory plasticity occurring specifically in the transition from prenatal to postnatal development.

Figures

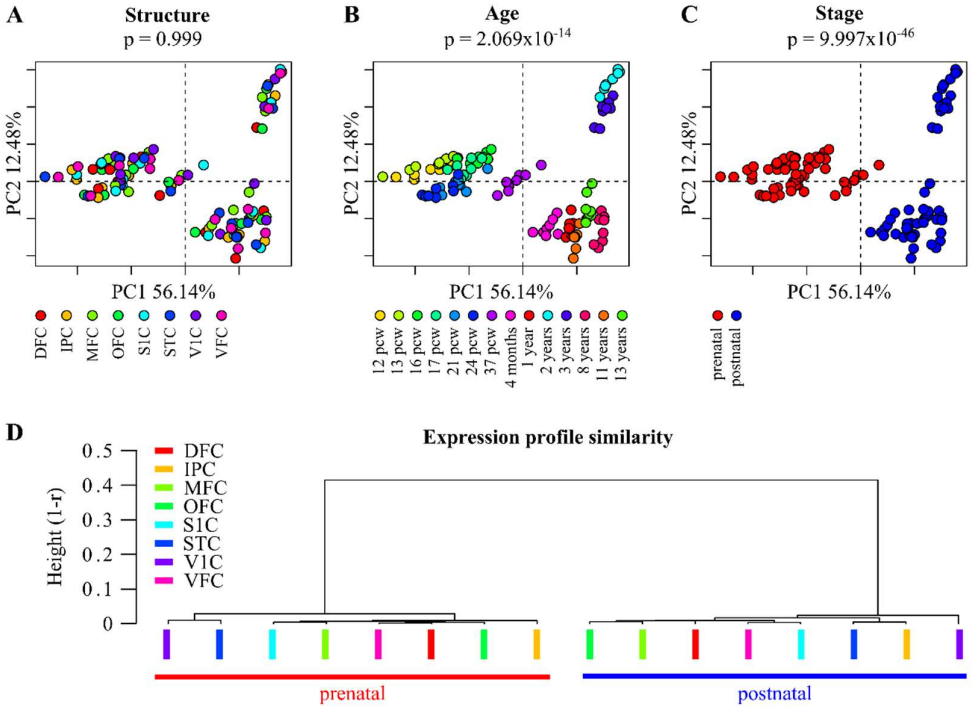


Figure 1. Developmental stage, but not anatomical structure contributes to the greatest component of variance in gene expression profiles. Multiple component analysis splitting samples of expression data by **A)** either structure, **B)** post conception age or **C)** prenatal/postnatal stage. Each plot shows the first and second components (together contributing to 68.83% of variance). Analysis of variance was carried out on PC1 to test for associations between this component and either, structure, post conception age or prenatal/postnatal stage. Associated p values are indicated. **D)** Relatedness between average postnatal or prenatal expression profiles across anatomical regions. Average expression per gene per cortical region was obtained for either prenatal or postnatal samples across all analysed cortical structures. Unsupervised hierarchical clustering was conducted using pairwise correlations between all resulting average expression profiles as a measure of similarity. Note, that the average expression profiles of any two prenatal regions are mores similar to each other, than they are to themselves across the perinatal boundary. Acronyms for brain structures: Dorsolateral prefrontal cortex (DFC), Posteroinferior parietal cortex (IPC), Medial prefrontal cortex (MFC), Orbital frontal cortex (OFC), Primary somatosensory cortex (S1C), Posterior superior temporal cortex (STC), Primary visual cortex (V1C) and Ventrolateral prefrontal cortex (VFC).

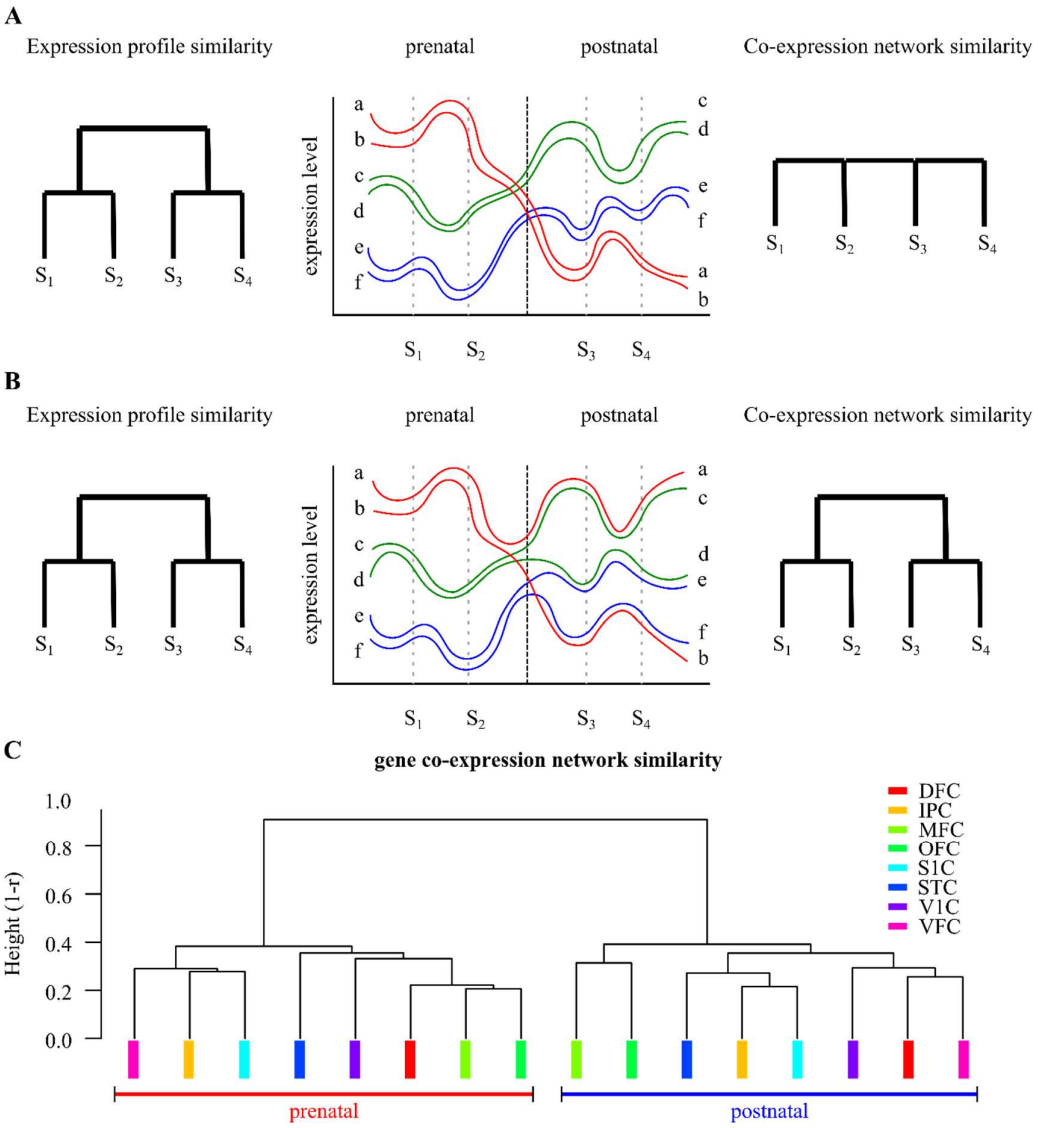


Figure 2. Schematic representation of two alternative models explaining the observed switch in the global expression profile sharply dividing the prenatal and postnatal brain development. Patterns of expression of six hypothetical genes in two prenatal (S1, S2) and two postnatal (S3, S4) samples. **A)** Switch in the global expression profile under constant regulatory interactions resulting from a pronounced change, during the perinatal boundary, in the overall level of expression of existing, but otherwise cohesive (constant coexpression structure), gene clusters. **B)** A similar switch in the global expression profile resulting from a widespread remodelling of the underlying regulatory structure leading to the reassembly of new functional clusters (varying coexpression structure). **C)** Relatedness between coexpression profiles across anatomical regions. Gene coexpression matrices per cortical region were obtained for both prenatal and postnatal samples. Unsupervised hierarchical clustering was conducted using pairwise correlations between al resulting coexpression

matrices as a measure of similarity. Note that the average coexpression structure of any two prenatal regions are mores similar to each other, than they are to themselves across the perinatal boundary.

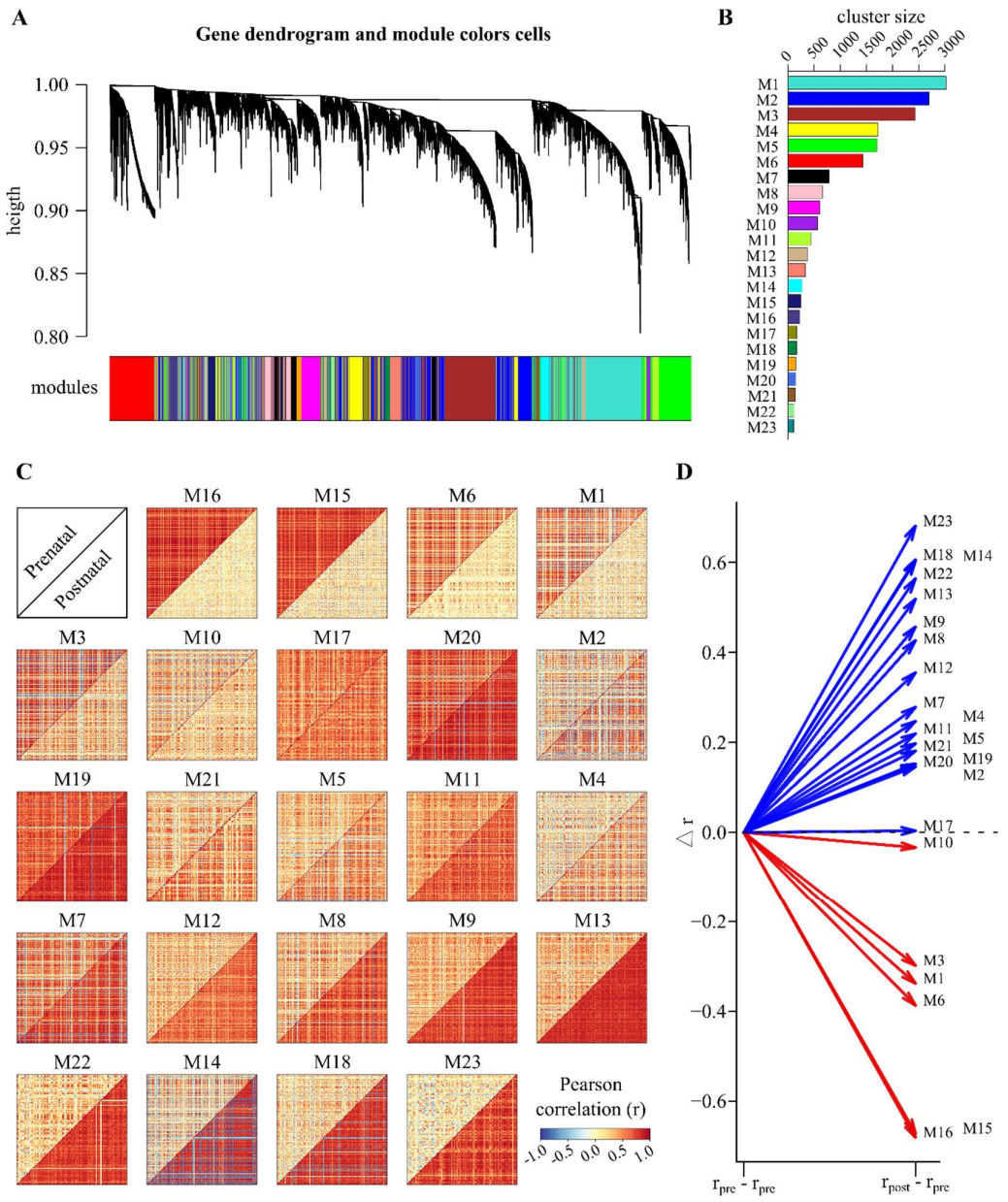


Figure 3. Differential coexpression network dendrogram and gene modules. A) Dendrogram showing average hierarchical clustering based on the topological overlap of the adjacency matrix defined by coexpression differences (see methods) and the corresponding gene modules indicated in different colours underneath. **B)** Histogram showing cluster size (number of genes) per module **C)** Heatmaps of the Pearson correlation coefficients between all possible gene pairs within each module. Each heatmap shows prenatal and postnatal coexpression separately (upper and lower diagonal respectively). Colour scale for correlation coefficients is shown at the top left corner of this panel. Modules are arranged from left to right and top to bottom according to the median correlation difference between prenatal and postnatal stage. **D)** Graph showing the change in the average correlation of each module

between pre and postnatal stages. Each arrow represents the difference in postnatal average correlation relative to the prenatal stage. 17 out of 23 differentially coexpressed modules displayed an increase in correlated expression in the postnatal stage with 6 modules showing reduced correlated activity in the same developmental stage.

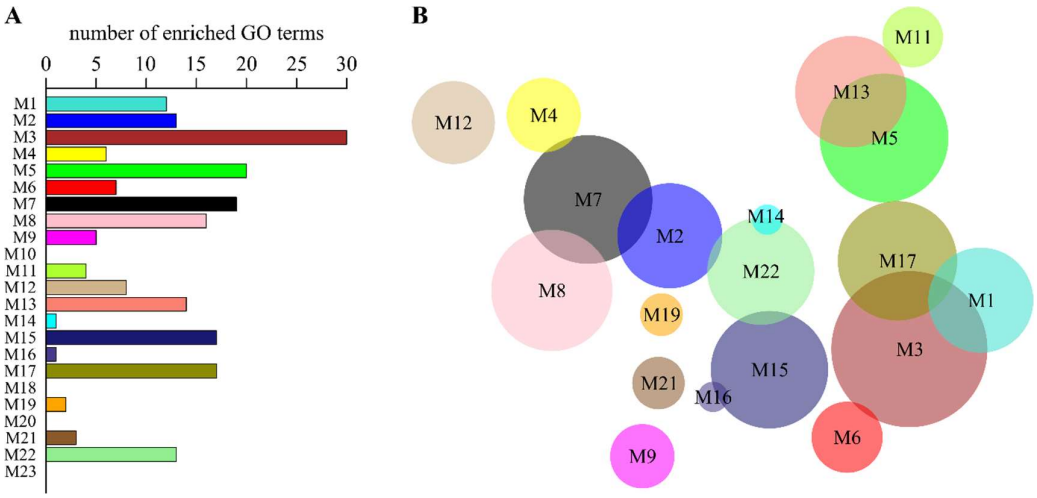


Figure 4. Regulatory reorganization modules are enriched in specific sets of biological processes. A) Barplot showing the number of significantly enriched GO terms per individual module. Statistical significance in enrichment of surveyed biological process GO terms (n = 106) was numerically assessed by comparing with 10 000 equally sized random samples of genes. Significance threshold (FDR < 0.05) was adjusted for multiple testing by Benjamini Hochberg correction. **B)** Venn-Euler diagram showing the relative number of gene ontology terms statistically overrepresented within each separate module. The area of each circle represents the number of enriched GO terms in each module and the overlap represents the relative proportion of overlapping GO terms between modules. Note that each module targets a quasi-exclusive set of biological functions with rare functional overlap between modules.

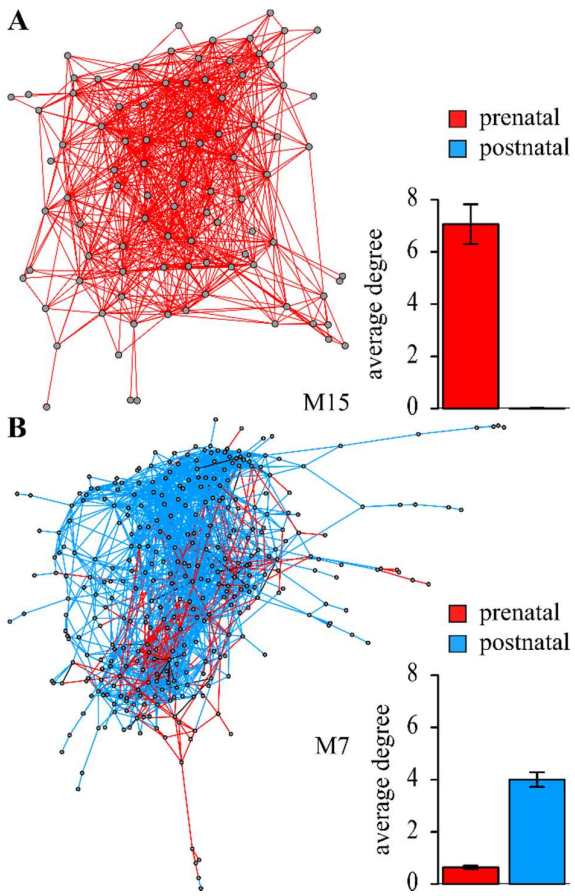


Figure 5. Network rearrangement at perinatal transition. Examples of the internal structure of the regulatory reorganization in two representative modules involving a combination of events of increased and decreased coordinated activity between individual gene pairs. Each module or gene cluster is represented as a graph of coexpression relationships, where nodes represent genes and edges represent high correlations ($r > 0.95$) between pairs of genes. Red edges represent prenatal-only high correlations, blue edges represent postnatal-only high correlations and black edges represent constant (prenatal and postnatal) high correlations. **A)** Module M15 shows an almost exclusive transition from low to high correlation in all involved genes in the transition from prenatal to postnatal development. **B)** By contrast, module M7 shows a transition from high to low coordination between genes in a subnetwork with a simultaneous transition from low to high correlated activity in a second sub-network. Inset: barplots showing the mean number of prenatal (red) or postnatal (blue) edges per node in each module.

Chapter 2.

Modular and coordinated expression of immune system regulatory and signalling components in the developing and adult nervous system

Jimena Monzón-Sandoval[1,2], Atahualpa Castillo-Morales[1,2], Sean Crampton[3], Laura McKelvey[3], Aoife Nolan[3], Gerard O'Keeffe[3,4]* and Humberto Gutierrez[1]*

[1] School of Life Sciences, University of Lincoln, Lincoln, UK

[2] Department of Biology and Biochemistry, University of Bath, Bath, UK

[3] Department of Anatomy and Neuroscience, Biosciences Institute, University College Cork, Cork, Ireland

[4] Irish Centre for Fetal and Neonatal Translational Research (INFANT), Cork University Maternity Hospital, Cork, Ireland

Abstract

During development, the nervous system (NS) is assembled and sculpted through a concerted series of neurodevelopmental events orchestrated by a complex genetic programme. While neural-specific gene expression plays a critical part in this process, in recent years, a number of immune-related signalling and regulatory components have also been shown to play key physiological roles in the developing and adult NS. While the involvement of individual immune-related signalling components in neural functions may reflect their ubiquitous character, it may also reflect a much wider, as yet undescribed, genetic network of immune–related molecules acting as an intrinsic component of the neural-specific regulatory machinery that ultimately shapes the NS. In order to gain insights into the scale and wider functional organization of immune-related genetic networks in the NS, we examined the large scale pattern of expression of these genes in the brain. Our results show a highly significant correlated expression and transcriptional clustering among immune-related genes in the developing and adult brain, and this correlation was the highest in the brain when compared to muscle, liver, kidney and endothelial cells. We experimentally tested the regulatory clustering of immune system (IS) genes by using microarray expression profiling in cultures of dissociated neurons stimulated with the pro-inflammatory cytokine TNF-alpha, and found a highly significant enrichment of immune system-related genes among the resulting differentially expressed genes. Our findings strongly suggest a coherent recruitment of entire immune-related genetic regulatory modules by the neural-specific genetic programme that shapes the NS.

Introduction

The orchestrated execution of the neural-specific genetic-code that governs the development and physiology of the nervous system (NS) is crucial for normal neural function (Lister et al 2013, Parikshak et al 2013, Willsey et al 2013). The importance of the correct execution of neural-specific gene expression for normal neural development and function, is highlighted by the fact that perturbations in the specific genetic networks that constitute this wider genetic code lead to neurodevelopmental and/or adult onset disorders of the NS (Brashear et al 2014, Helsmoortel et al 2014).

Intriguingly, disturbances in the immune system (IS) is a common theme in a wide variety of disorders of the NS, ranging from childhood autism to depression (Mitchell & Goldstein 2014). While this could result from direct immune responses (mainly inflammatory) disrupting the neurodevelopmental programme, neurological disorders triggered by the IS could also result from a potential genetic regulatory overlap between the IS and the NS. In support of this possibility, in recent years, a number of signalling molecules and regulatory components originally described in the IS have been found involved in distinct neural-specific functions including early survival of neuronal precursors, dendritic and axonal growth in developing neurons as well as synaptic remodelling and learning and memory mechanisms both in the developing and adult NS (Carriba et al 2015, Galenkamp et al 2015, Gavalda et al 2009, Gutierrez & Davies 2011, Gutierrez et al 2005, Gutierrez et al 2013, Nolan et al 2011, O'Keeffe et al 2008, Twohig et al 2011). Furthermore, immune system-related genes have been found statistically overrepresented among highly variable genes expressed during early brain development (Sterner et al 2012) suggesting a critical role for these genes at key stages of NS development. More broadly, changes in gene family size associated with increased encephalization in mammals have been found enriched in IS related functions prominently expressed in the NS (Castillo-Morales et al 2014). Taken together these findings suggest a potentially wider involvement of large numbers of immune system-related genes in key aspects of NS development and function.

While the involvement of isolated immune-related signalling components in neural functions may reflect their otherwise ubiquitous character, it may also reflect a much wider, as yet undescribed, genetic network of immune–related molecules acting as an intrinsic component of the neural-specific regulatory machinery that shapes the functional complexity of the NS. In order to gain insights into the scale and wider functional organization of immune-related genetic networks in the developing and adult NS, we examined the large scale pattern of expression of these genes in the developing and adult brain. A complex phenotype is usually the result of an assembly of molecular and genetic components acting in concert (Hartwell et al 1999). As a result, genes involved in related cellular responses display coordinated pattern of expression reflecting their functional association (Eisen et al 1998, Homouz & Kudlicki 2013, Obayashi & Kinoshita 2011, Saris et al 2009, Torkamani et al 2010, Zhang et al 2012, Zhang 2012). In the NS, this functional organization of closely coordinated genes also displays a substantial degree of conservation across species (Oldham et al 2006, Oldham et al 2008).

In this study, because of their conspicuous presence and involvement in neural-specific functions, we specifically asked whether immune-related regulatory and signalling components operate in isolation from each other but in close association with neural-specific genes or, alternatively, in closer coordination with other immune-related genes.

By combining human gene expression data analysis, with microarray profiling in dissociated developing neurons, this work provides compelling evidence showing that immune-related genetic networks form an intrinsic part of the wider genetic programme that governs NS development and function. We discuss the implications of these findings and their potential relationship to disorders of the NS.

Materials and Methods

Gene Expression Data

We used RNA-seq expression data the from the Allen's Institute Brainspan database (http://www.brainspan.org/). Reads per kilobase of transcript per million reads

mapped (RPKM)-normalized data in this dataset were summarized to Ensembl Gene IDs, and further normalized against total expression per sample. In order to ensure a homogenous representation of structures at all development points used, we selected a subset of this dataset covering 12 brain regions (A1C, CB, DFC, IPC, ITC, M1C, MFC, OFC, S1C, STC, V1C, VFC) and 20 developmental time points ranging from 12 post conception week through to 40 years (12, 16, 21 and 24 pcw; 4 and 6 months; and 1, 2, 8, 11, 13, 15, 18, 19, 21, 23, 30, 36, 37, 40 years old).

Human microarray expression data for normal brain (GSE13162), muscle (GSE11681), kidney (GSE2004), liver (GSE2004) and aortic endothelium (GSE29903) was obtained from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), only samples categorized as "normal", "control" or "healthy" were considered for analysis. Either RMA or median normalized values were summarized to Ensembl Gene IDs by averaging probe expression as described above. Probes that matched multiple Ensembl Gene IDs were excluded from the analysis.

Gene Ontology Annotations

The lists of genes annotated to the "immune system process" (IS, Gene ontology ID: GO:0002376) and "Neurological system process" (NS, GO:0050877) were obtained from the GO slim GOA database (http://geneontology.org/) through Ensembl Biomart (http://www.ensembl.org/biomart/). Genes annotated to both categories were excluded from all the analysis, so as to not overestimate the relation between IS and NS. For the functional enrichment analysis in "IS process" sub-categories (Table 2), we used the annotation available in Ensembl version 73 and used OBOedit software to assign all children GO categories to their corresponding GO parent.

Co-Expression and Clustering Analyses

Co-expression analyses were carried out by obtaining the Pearson correlation coefficient across all possible pairs of IS-associated or IS-NS-associated genes in the brain, as well as co-expression within IS genes in other human tissues. Measurements of clustering coefficients and statistical analyses including numerical simulations were carried out using the R statistical software package, and particularly *igraph* library.

Functional Enrichment Analysis

Over-representation of IS process genes in our microarray data was measured by contrasting the number of genes annotated to a relevant GO with the expected representations of GO terms, with their standard deviations numerically derived from Monte Carlo simulations using at least 10,000 equally-sized random samples of genes from the list of Ensembl Gene IDs covered by the microarray profiling (n = 15,888). Benjamini-Hochberg multiple testing adjustments against the number of IS process sub-categories tested (n = 21) were carried out. Categories with a resulting adjusted p-value < 0.05 were deemed significantly enriched.

Cell Cultures

The superior cervical ganglia (SCG) of embryonic and postnatal Sprague-Dawley rats (Biological Services Unit, UCC) were dissected out and grown in Dulbecco's Modified Eagle Medium/F12 (DMEM:F12, Sigma) containing 1% penicillin/streptomycin (Sigma), 1% glutamine (Sigma), 1% N2 (Invitrogen), 2% B27 (Invitrogen) and 10 ng/ml NGF (R&D Systems). Dissociated neurons were plated at a low density on poly-ornithine/laminin-coated 4 well 35 mm tissue culture dishes (Sigma; Greiner Bio-One). TNF-alpha (10 ng/ml, Promokine) was added to the medium and the cells were incubated under a humidified atmosphere containing 5% CO2 at 37°C for 24 h.

Microarray Profiling

After 24 h in culture, medium was removed and the cells were washed twice in PBS and total RNA was isolated using the RNeasy mini extraction kit (Qiagen) according to the manufacturer's instructions. The samples were run in an AGILENT bioanalyser to check RNA quality/integrity prior to being sent for microarray analysis. Microarray hybridization was outsourced through a commercial provider (Source Biosciences) using the Affymetrix GeneChip Rat 1.0 ST array with a 2 ug RNA analysed per group. Array images were reduced to intensity values for each probe using Affymetrix MAS 5.0 software. RMA was used to correct background and normalize probe levels (AffyPackage). Probes with expression values lower than the average of negative controls in every sample was removed. Expression values were summarized to

Ensembl Gene IDs by averaging probe expression. Finally, gene expression was normalized against the total signal level in each sample.

Results

Highly Correlated Expression of Immune System Genes in the Nervous System

In order to characterize the functional interactions of IS genes in the NS; we used gene co-expression network analysis, an approach that has been widely used to gain insights into the functional organization of transcriptomes across tissues, conditions and species (Eisen et al 1998, Homouz & Kudlicki 2013, Obayashi & Kinoshita 2011, Saris et al 2009, Torkamani et al 2010, Zhang et al 2012). We started by asking whether IS genes show a stronger pattern of coordinated activity with genes specifically involved in neural functions relative to the background gene population. To this end, we used gene expression data obtained from the normal developing human brain (Hawrylycz et al 2012) and obtained the list of genes annotated to the "Immune system process" and "Neurological system process" from the GO slim GOA database.6 In order to focus our study on those genes separately involved in these two processes, we excluded genes simultaneously annotated to both categories. This resulted in a total of 1584 and 1036 genes annotated to "Immune system process" and "Neurological system process", respectively, for which human expression data was available in the Brainspan dataset.

To quantify the overall co-expression within and between these groups of genes, we obtained the median Pearson correlation coefficient of all possible pairs of IS and NS genes as well as all possible pairs of IS-IS genes and contrasted these values with the expected median correlation derived from 10,000 equally sized samples of random genes drawn from the background gene population. As shown in Figure 1, the highest level of co-expression in the brain occurs between IS genes and other IS genes (IS-IS interactions, $p = 0.003$) followed in second place by the median co-expression between IS and NS genes (IS-NS interactions, $p = 0.018$). It is noteworthy that NS-NS co-expression is also remarkably high when compared to equally-sized random samples (Z-score 4.45, $p < 0.0001$ not shown in figure). By contrast the median correlation between IS genes and any random non-specific genes on the other hand was not

significantly different from chance expectations ($p > 0.1$). These results demonstrate that, in the NS, IS genes are more highly co-expressed with other IS genes than expected by chance and that this level of co-expression was also higher than the observed co-expression between IS genes and genes associated with neurological system processes.

Highly Correlated Expression is not a General Feature of Immune System-Related Genes in Non-Nervous Tissues

If the highly coordinated expression of IS genes observed in nervous tissue was also found outside the NS, it could reflect the ubiquitous modular organization of IS genes and not their particular recruitment by the NS. We tested this hypothesis by analysing independent microarray gene expression data for various human tissues. To this end, we used data obtained with the same microarray platform and derived from human tissues for which at least six biological samples were available. We found five datasets meeting these requirements corresponding to whole brain, liver, aortic endothelium, muscle and kidney (see "Materials and Methods" Section).

After normalization, we obtained for each tissue the median co-expression between all possible pairs of IS genes and compared these values to the expected median co-expression obtained from 10,000 equally sized random samples of background genes. As shown in Figure 2, the highest level of correlated expression of IS-related genes was found in whole brain ($p < 10^{-16}$), further confirming our previously observed high coordination of IS genes in the NS. To a much lesser extent, muscle tissue also displayed a high coordination of IS-related genes. By contrast, liver, kidney and endothelium showed non-significant levels of coexpression when compared with random expectations ($p > 0.1$). These results demonstrate that the observed highly concerted expression of IS-related genes in the NS is not a general feature of the expression pattern of these genes across tissues.

Immune System-Related Genes Display a High Regulatory Clustering in the Developing and Adult Nervous System

The observed bias in the median correlation of IS genes in the NS could be the result of either a large number of isolated pairs of highly correlated genes or, alternatively, the result of a large highly interconnected network of tightly co-regulated genes.

In order to obtain information on the structure of the involved interactions, we focused on all existing correlations above a given threshold ($R > 0.9$). The resulting map of interactions can be represented as a network, where nodes represent genes and edges represent existing correlations with a coefficient value above 0.9. At this threshold, 662 IS genes were found involved in strong correlations either with other IS genes or any other genes. Given a set number of interactions (edges), the resulting network can display different degrees of clustering between existing nodes. In order to assess the degree of clustering, we obtained the clustering coefficient of each node. This index quantifies, for a given node, the fraction of existing interactions between all immediate neighbours (Watts & Strogatz 1998). Next, we compared the average clustering coefficient of the whole IS-IS network with that of a simulated random network where the number of nodes, number of edges and distribution of edges per node (degree distribution) was the same as the real network. As shown in Figure 3A the observed mean clustering coefficient of the network of highly correlated IS genes was considerably higher than that of the random control network. In order to estimate the significance of this bias, we compared the mean clustering coefficient of IS genes, with the expected mean clustering coefficient derived from 10,000 simulations of equally sized random networks with the same number edges and degree distribution (edges per node). As shown in Figure 3B, the observed clustering coefficient was on average five times higher than the average expected value ($p < 0.0001$). The observed high mean clustering coefficient of IS genes relative to chance expectations holds over a range of correlation cut-off values as shown in Figure 3C, overall demonstrating that IS genes display a significantly high level of clustered co-expression with other IS genes in the NS. It is also worth noting that within the network of highly correlated IS genes, the overall proportion of strong correlations among IS genes (number of IS-IS edges) is also significantly higher than expected by chance relative to the total number of associations involving IS genes ($X^2 = 8.912$, $p = 0.0028$).

Coordinated Activation of Large Numbers of Immune System Genes in Primary Cultured Neurons

Taken together, the above results demonstrate that, in the NS, IS-related genes display a stronger overall association with other IS genes than with genes involved in neural-specific functions. In addition, IS genes display a higher average clustering than expected by chance further revealing that IS-related signalling and regulatory components operate under a tightly coordinated pattern of transcriptional regulation in the NS when compared with background genes. Based on these findings, we hypothesized that upon stimulation, IS genes will trigger the simultaneous activation of a large number of other IS-associated genes in isolated neurons.

We tested this prediction using gene expression microarray profiling in dissociated cultures of developing sympathetic neurons derived from the SCG. We chose this neuronal population for three reasons: (1) SCG neuronal cultures offer an experimentally tractable model consisting in >95% neurons from a homogeneous neuronal population, thereby eliminating the confounding influence of highly heterogeneous populations of neurons and glia present in dissociated cultures from other regions of the NS (Glebova & Ginty 2005, Orike et al 2001); (2) These neurons have been shown to respond to tumor necrosis factor alpha (TNF-alpha), a cytokine that negatively regulates axonal growth during early postnatal stages (Kisiswa et al 2013, Nolan et al 2014, Twohig et al 2011), thereby offering an ideal opportunity to experimentally assess the effect of an immune-associated cytokine on the global gene expression of dissociated neurons; and (3) While central and peripheral neural tissues share over 99% of their gene expression profiles (LeDoux et al 2006, Smith et al 2011), the use of peripheral neurons provides an additional opportunity to test whether the observed regulatory clustering of immune related genes extends to both central and peripheral neurons.

Accordingly, we stimulated cultured neonatal SCG neurons with TNF alpha (10 ng/ml) for 24 h and cells were then collected for total RNA extraction and preparation for gene expression profiling (see “Materials and Methods” Section).

We identified differentially expressed genes based on the rank products method (Hong et al 2006) and measured the overall response of IS-associated genes by analysing

enrichment of IS-associated genes among those genes displaying the highest levels of differential expression (upregulation) upon TNF-alpha stimulation ($p < 0.05$, $n = 302$).

As shown in Table 1, significantly up-regulated genes showed a statistically significant enrichment of IS process-annotated genes. In order to identify subcategories of IS genes particularly over-represented among these genes, we measured enrichment of all gene ontology subcategories associated with IS process relative to the background gene population. This analysis revealed a significant over-representation of genes associated with immune response, leukocyte migration, regulation of IS process, positive regulation of IS process and activation of immune response (Tables 2, 3). These results demonstrate that, in isolated neurons, upon stimulation, IS genes trigger the simultaneous activation of a disproportionally large number of other IS-associated genes when compared to chance expectations.

Discussion

A wide variety of homeostatic perturbations of the NS including pathogen invasion, endogenous disease and injury, are known to induce an inflammatory response which often involves infiltration of immune cells and activation of resident effectors, such as microglia. While the complex interaction between the immune surveillance machinery and the NS has been the focus of a large number of studies in the past (Ousman & Kubes 2012, Ransohoff & Brown 2012), in recent years a growing number of signalling and regulatory components of the IS have emerged as key molecular players in a wide variety of neural-specific functions ranging from early survival of neuronal precursors and dendritic and axonal growth in developing neurons through to synaptic remodelling and learning and memory mechanisms (Carriba et al 2015, Galenkamp et al 2015, Gavalda et al 2009, Gutierrez & Davies 2011, Gutierrez et al 2005, Gutierrez et al 2013, Nolan et al 2011, O'Keeffe et al 2008). Whether these findings reflect a wider and generalized involvement of IS-related regulatory and signalling components in neural-specific mechanisms, is unknown.

By interrogating the wider regulatory organization of IS-associated signalling and regulatory components in the developing and adult NS, in this study we have found that IS genes are more highly co-expressed with other IS genes than expected by

chance. In addition, we found that the underlying co-expression network of highly associated IS genes displays a much higher clustering coefficient than expected in networks with equal density and degree distribution. These results reveal a strong underlying regulatory association between large numbers of IS genes operating in the NS.

The recruitment, in the NS, of signalling and regulatory components of the IS could, in principle, simply reflect the ubiquitous character of these regulators, and we would therefore expect them to establish, close but independent functional associations with the wider machinery of neural-specific functions. However, in the NS, IS genes display a stronger association and regulatory clustering with other IS genes than with the wider molecular machinery involved in neural-specific functions suggesting a coherent functional recruitment of entire segments of the IS regulatory machinery by the NS.

We experimentally tested this regulatory clustering in dissociated neurons by conducting microarray gene expression profiling of isolated developing neurons in culture, and found that stimulation with the pro-inflammatory cytokine TNF-alpha triggers the simultaneous activation of a disproportionally large number of other IS-associated genes. The fact that we tested a prediction derived from human expression patterns in experimentally tractable cultured neurons derived from the developing rat, further suggests that the coherent recruitment of IS regulatory clusters is conserved between rodents and humans. While subtle differences at this level may exist between the rodent and human model, this finding is in line with the fact that most basic aspects of extra and intracellular signalling events and their underlying networks of regulatory interactions show a remarkable degree of conservation between murine and human cell models (Herschkowitz et al 2007, Shortman & Liu 2002). In addition, the fact that our experimental test was carried out in peripheral neurons further shows that the observed regulatory clustering of immune related genes extends to both central and peripheral NS, an observation otherwise consistent with the highly similar gene expression profiles reported for central and peripheral neurons (LeDoux et al 2006, Smith et al 2011).

In addition, by comparing patterns of coordinated expression of IS-related genes outside the NS, including cell types where immune regulators are known to play

central roles such as hepatocytes and endothelial cells (Gargalovic et al 2006), we demonstrate that the highly coherent expression of IS-associated genes observed in the NS is by no means a general feature of these genes outside neural tissues.

Taken together, our results support the notion of a widespread and modular recruitment of IS regulatory and signalling circuits by the NS developmental programme in mammals.

Given the tight regulatory association of large numbers of IS genes in the developing and adult NS, our results raise the possibility of numerous instances of potential interference with neural physiology arising from organismal immune states not directly related with immune surveillance or inflammation in the NS.

Thus, for instance, during pregnancy, maternal infection in the second trimester increases the risk, for affected offspring, of developing psychiatric and neurological disorders such as schizophrenia and autism (Atladottir et al 2010, Boksa 2010, Sorensen et al 2009). The mechanisms linking maternal inflammation with defective neural development however are unclear. Furthermore, maternal infection in rodents during late gestation results in morphological, electrophysiological and molecular changes in the brains of offspring (Garbett et al 2012). While this could result from direct immune responses (mainly inflammatory) disrupting the neurodevelopmental programme, our findings suggest that neurological disorders triggered by the IS could also result from the underlying genetic regulatory overlap between the immune and the NS. Interestingly, maternal infection in rodents also triggers changes in proinflammatory cytokine levels in the fetal brain and fetal blood (Garay et al 2013). Whether these changes can interfere with the establishment of normal developmental patterns in neurons is unknown. However, our results would predict that systemic changes in pro-inflammatory cytokines could potentially trigger concomitant expression changes in IS-related genes in developing neurons leading to measurable alterations in their developmental programme.

Conclusion

In summary, our results demonstrate that IS genes display a significantly strong level of concerted regulation and transcriptional clustering in the developing and adult brain, supporting the notion of a coherent and widespread recruitment of IS regulatory components by the NS developmental programme in mammals. These results further provide a genetic basis for potential interference with neural functions arising from systemic changes in immune surveillance and inflammatory states.

Tables

Table 1. GO enrichment analysis of differentially expressed (up regulated) genes relative to background gene population.

GO Term ID	Category name	Observed genes	Expected genes	Numeric *p* value
GO:0002376	immune system process	22	8.9139	< 0.0001

Table 2. GO enrichment analysis within immune system (IS) process genes relative to background gene population.

GO Term ID	Category name	Observed genes	Expected genes	Numeric *p* value	Adjusted numeric *p* value
GO:0006955	immune response	22	7.2773	0	0
GO:0050900	leukocyte migration	9	2.0063	0	0
GO:0002682	regulation of immune system process	19	8.1524	2×10^{-4}	0.001
GO:0002684	positive regulation of immune system process	13	5.0599	5×10^{-4}	0.002
GO:0002253	activation of immune response	6	2.2173	0.0067	0.022333

Table 3. List of IS-associated genes significantly up regulated in response to TNF-alpha stimulation of developing sympathetic neurons.

Ensembl Gene ID	RGD Symbol	Description	immune system process	immune response	leukocyte migration	regulation of immune system process	positive regulation of immune system	activation of immune response
ENSRNOG00000007917	Cd46	CD46 molecule, complement regulatory protein		1		1	1	1
ENSRNOG00000011971	C1s	complement component 1, s subcomponent	1	1		1	1	1
ENSRNOG00000014832	Mapkapk3	mitogen-activated protein kinase-activated protein kinase 3	1	1		1	1	1
ENSRNOG00000019440	Kcnn4	potassium intermediate/small conductance calcium-activated				1	1	1
ENSRNOG00000033134	Mef2c	myocyte enhancer factor 2C		1		1	1	1
ENSRNOG00000033879	Clec7a	C-type lectin domain family 7, member A		1		1	1	1
ENSRNOG00000000239	Cdl7	chemokine (C-C motif) ligand 7	1	1	1	1	1	
ENSRNOG00000004498	Scin	scinderin				1		
ENSRNOG00000008409	Myo1f	myosin IF	1	1		1		
ENSRNOG00000009912	Fgr	FGR proto-oncogene, Src family tyrosine kinase	1	1		1	1	
ENSRNOG00000010906	Cdl5	chemokine (C-C motif) ligand 5	1	1	1	1	1	
ENSRNOG00000013794	Rbp1	retinol binding protein 1, cellular				1		
ENSRNOG00000014333	Vcam1	vascular cell adhesion molecule 1	1		1	1	1	
ENSRNOG00000015618	Wnt5a	wingless-type MMTV integration site family, member 5A		1	1	1	1	
ENSRNOG00000016294	Cd4	Cd4 molecule	1	1		1	1	
ENSRNOG00000018659	Csf1	colony stimulating factor 1 (macrophage)	1	1		1		
ENSRNOG00000024899	Cxcl13	chemokine (C-X-C motif) ligand 13	1	1	1	1	1	
ENSRNOG00000028015	Pf4	platelet factor 4	1	1	1	1		
ENSRNOG00000032224	Hist2h4	histone cluster 2, H4				1		
ENSRNOG00000008837	Ass1	argininosuccinate synthase 1	1	1				
ENSRNOG00000016535	Cd22	chemokine (C-C motif) ligand 22	1	1				
ENSRNOG00000022298	Cxcl11	chemokine (C-X-C motif) ligand 11	1	1				
ENSRNOG00000026647	Cxcl16	chemokine (C-X-C motif) ligand 16	1	1	1			
ENSRNOG00000028548	Cdl9	chemokine (C-C motif) ligand 9	1	1				
ENSRNOG00000028768	LOC100911495	guanylate-binding protein 4-like	1	1				
ENSRNOG00000031743	Gbp2	guanylate binding protein 2, interferon-inducible		1				
ENSRNOG00000032240	Gbp5	guanylate binding protein 5	1	1				
ENSRNOG00000017197	Pdgfb	platelet-derived growth factor beta polypeptide	1		1			
ENSRNOG00000043451	Spp1	secreted phosphoprotein 1	1		1			
ENSRNOG00000011238	Tinarp	TCDD-inducible poly(ADP-ribose) polymerase	1					
ENSRNOG00000019494	Psmb10	proteasome (prosome, macropain) subunit, beta type 10	1					

Figures

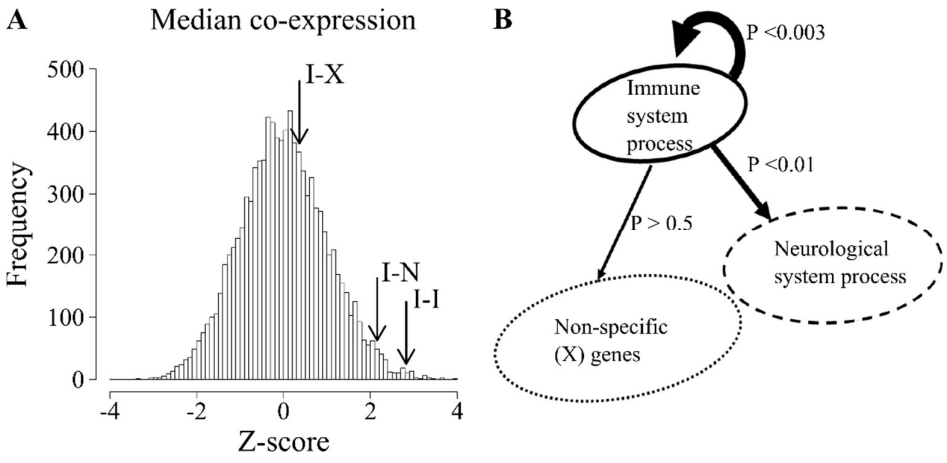


Figure 1. Immune system (IS) process associated genes are highly co-expressed in the developing and adult brain. A) Distribution of median co-expression values of 10,000 equally sized random samples of genes. Co-expression is expressed as Z score-transformed median correlation coefficients relative to the expected distribution. Arrows show the observed median co-expression between IS process genes (I-I), IS process and neurological system process genes (I-N) as well as IS process genes and random non-specific genes (I-X). **B)** Schematic representation of the statistical bias (p values) in co-expression between the indicated populations of genes.

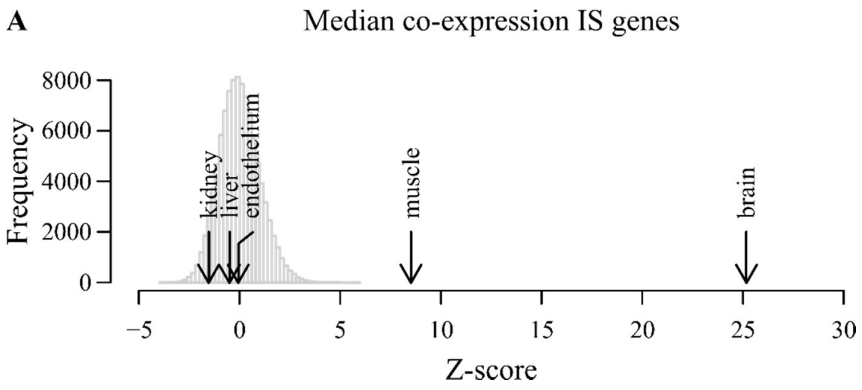


Figure 2. Correlated expression is not a general feature of immune system-related genes in non-nervous tissues. Co-expression of IS genes was examined in independent microarray expression data derived from whole brain as well as liver, endothelial cells, kidneys and muscle. Histogram shows the distribution of median co-expression values of 10,000 equally-sized random samples of genes in five different tissues. All expected and observed co-expression values and corresponding distributions were Z score-transformed in order to compare across tissues. Arrows show the observed Z-score transformed median co-expression between IS process genes in the indicated tissues.

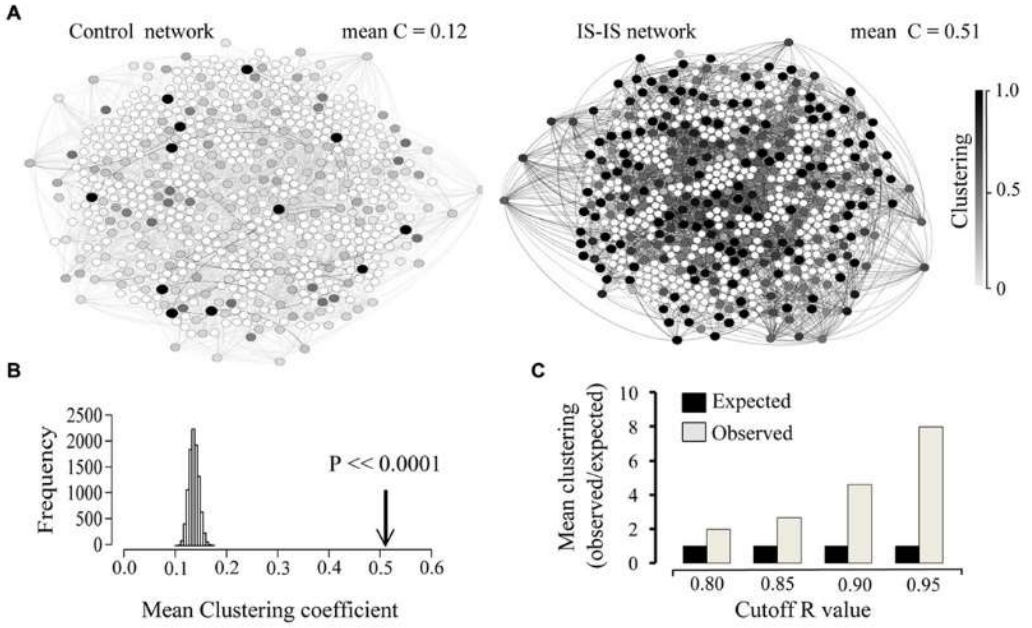
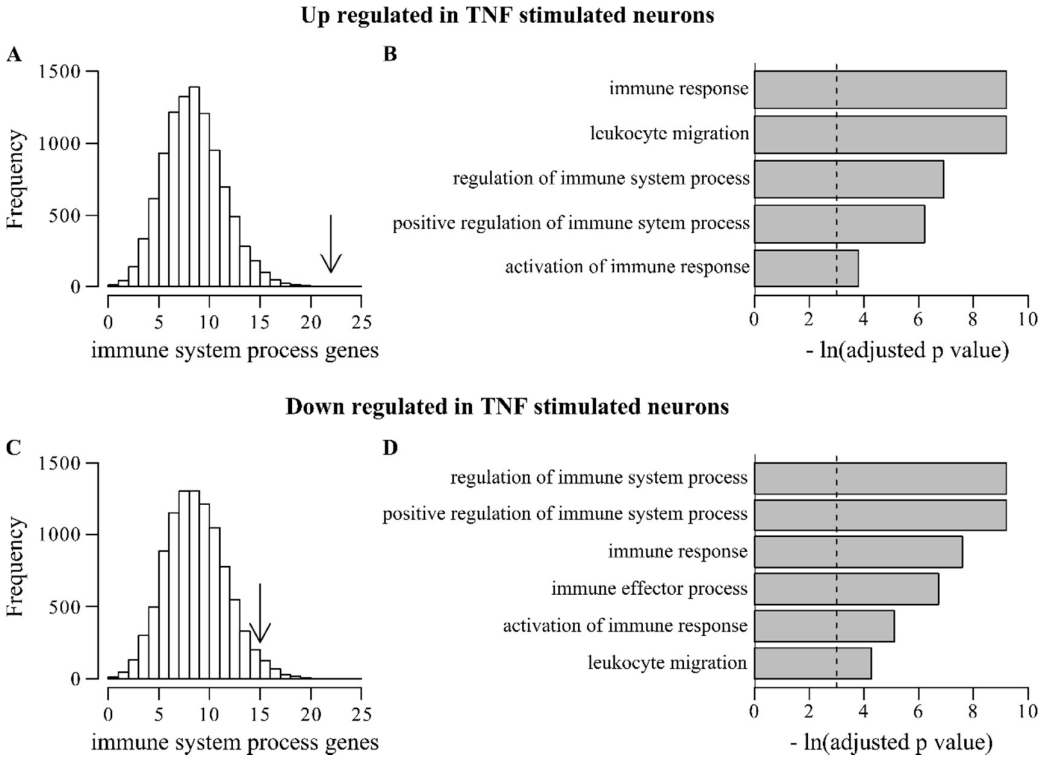


Figure 3. Immune system genes display a high co-expression clustering in the developing and adult brain. A) Network representation of 662 IS genes strongly correlated (edges = r > 0.9) with other genes and an equivalent random network with the same size (number of nodes), density (number of edges) and degree distribution (edges per node). Only associations between IS genes are shown (IS-IS links). The clustering coefficient of each node is represented by the corresponding colour intensity (white = 0, black = 1). Mean clustering coefficient of each network is indicated. **B)** Comparison between the observed Clustering (arrow) and the distribution of expected mean clustering coefficients obtained from 10,000 simulated random networks with same size, density and degree distribution as the real network. **C)** Chart showing the ratio of observed vs. expected clustering coefficient across a range of correlation coefficient value cut-off thresholds.

Supplementary figure



Supplementary Figure 1. Up and down regulated genes in TNF alpha stimulated SCG rat neurons are enriched in Immune system process genes. Enrichment analysis for immune system process genes was assessed for gene up and down regulated in TNF stimulated neurons. Distributions in **A)** and **C)** show the expected number of immune system genes obtained from a Monte Carlo simulation, the observed number of genes is represented by an arrow. **B)** and **D)** show the overrepresentation of specific IS process child terms, where bars denote Benjamini-Hochberg adjusted numeric p value, dashed line represents a significant level ($p_{BH\ adjusted} < 0.05$) were the overrepresentation was considered significant.

Chapter 3.

Dissecting yap1 function through coexpression analysis in the human brain.

Jimena Monzón-Sandoval[1, 2], Atahualpa Castillo-Morales[1, 2], Humberto Gutierrez[2], Makoto Furutani-Seiki[1], Araxi O. Urrutia[1]

[1] Department of Biology and Biochemistry, University of Bath, Bath, UK

[2] School of Life Sciences, University of Lincoln, Lincoln, UK

Abstract

Involved in the regulation of organ size, cell growth, proliferation, apoptosis, survival, differentiation and tissue tension, *yap1* is a key regulator in the cell. Here we apply a coexpression approach to identify probable *yap1* interactors in a non-proliferative context by contrasting two periods of human brain development that are very different in their rates of proliferation. We found a mostly different set of genes coexpressed to *yap1* in postnatal brain development to those coexpressed in an earlier stage. By a couple of complementary approached we identify that the observed difference in the sets of coexpressed genes is not expected by chance. Moreover, we found a marked difference in the cellular component associated to each group of *yap1* coexpressed genes, as well as distinct biological processes and molecular functions overrepresented at the *yap1* coexpressed genes of each developmental stage. An enrichment analysis of microRNAs and transcription factor binding sites pointed towards a potentially dissimilar regulation of *yap1* pontential interactors. Furthermore, known protein-protein interactors of YAP1 were found overrepresented among *yap1* coexpressed genes as well as down regulated genes in a *yap1* knockdown in human spheroids. Our results allow us to identify close functional associates of *yap1*during this important developmental transition, unveiling new roles in non-proliferative functions.

Introduction

YAP1 (Yes associated protein 1) has been classically studied as a downstream effector of the Hippo signalling pathway. Briefly, when the Hippo pathway is activated, MTS1 and MTS2 are in a phosphorylated state which lead to the phosphorylation of LATS1 or LATS2 which directly bind and phosphorylates YAP1, which is retained in the cytoplasm by 14-3-3 and catenins (Low et al 2014). In turn, when the Hippo pathway is repressed hipophosphorylated YAP1 can translocate to the nucleus and bind typically to TEAD transcription factors promoting gene expression of a large number of target genes (Zhao et al 2008). Regulation of organ size, cell growth, proliferation, survival and differentiation are among the functions where YAP1 plays a role as the main effector of the Hippo signalling pathway (Camargo et al 2007, Lian et al 2010).

Although most strongly associated as a member of the Hippo pathway driving cell proliferation, YAP1 has also been associated with other pathways. In addition to its role in the Hippo signalling pathway, YAP1 has been involved in other pathways such as Notch, EGF, TGFβ, and Wnt (Attisano & Wrana 2013, Fan et al 2013, Li et al 2012, Wang et al 2014). The activation of these pathways may depend on different cellular cues such as cell-cell contact, adhesion cues, cell polarity, extracellular signals, mechanotransduction and cellular stress (Low et al 2014, Zhao et al 2012). For example, YAP1 translocation from the cytoplasm to the nucleus has been associated to cell density in cellular cultures, at higher density YAP1 is retained in the cytoplasm while at lower density it can translocate to the nucleus where it can promote the transcription of target genes along with TEAD (Zhao et al 2007). More recently, a novel function of YAP1 controlling the regulation of tissue tension in medaka and zebra fish has been described (Porazinski et al 2015).

Studies have revealed that YAP1 plays key roles during development both in driving cell proliferation but it also has other non-proliferative functions mediated by the interactions of YAP1 with a large number of proteins. YAP1 gene contains two WW domains consistent of two tryptophans (W) spaced by approximately 20 amino acids, these domains are quite versatile and allow YAP1 to preferentially bind different interactors (Iglesias-Bexiga et al 2015). Two main isoforms can be derived from YAP1 by alternative splicing, the first isoform contain only one WW domain while the

second contains two WW domains. Both isoforms also contain a SH3-binding motif and a Transcriptional Activating Domain (TAD), plus a PDZ binding motif necessary for the nuclear translocation of YAP1 (Sudol et al 2012). A number of proteins can bind to YAP1 at its different domains, for example, AMOTL and LATS1 have been shown to bind YAP1's WW domain (Oka et al 2008, Zhao et al 2011). In addition to the four TEAD transcription factors YAP1 has been found to interact either directly or indirectly with other transcription factors such as E2F, P73, SMADs, TBX5, CTNNB, ERB4 and RUNX2 (Alarcon et al 2009, Ehmer et al 2014, Kapoor et al 2014, Komuro et al 2003, Levy et al 2007, Rosenbluh et al 2012, Strano et al 2001, Zaidi et al 2004) potentially targeting the transcription of different sets of genes. Over a hundred proteins have been shown to directly interact with YAP1 in HEK293 cells (Hergovich 2012) providing an indication of how complex and context dependent is the molecular network of YAP1. Despite the relatively well known interaction network of YAP1 as part of the Hippo pathway, the gene interaction network of YAP1 in non-proliferative functions in particular as a mediator of tissue tension remain poorly understood.

Coexpression analyses provide a measure of the likely association between genes. It has been shown that a pair of highly coexpressed genes have a better chance to participate in the same biological function (Allocco et al 2004). Coexpression networks have also been widely used to annotate uncharacterized or poorly characterized genes (Wolfe et al 2005).

Comparison of coexpression networks even from closely related species have uncovered numerous differences in gene interactions which might underlie phenotypic differences (Oldham et al 2006, Stuart et al 2003). Studying coexpression networks in different conditions can reveal changes in the gene regulatory architecture. Such an approach has been used in the context of disease in which a group of genes may change its co-expression at the disease state when compared to a healthy state (Choi et al 2005, de la Fuente 2010, Miller et al 2008). Even in normal conditions, there has been shown that some genes may decrease their coexpression in mice as they age (Southworth et al 2009). This coexpression changes may indicate a rearrangement of the network in response to diverse stimuli at different levels (Tesson et al 2010).

During normal human brain development the most prominent processes involve cell proliferation, migration, synaptogenesis and myelination. The main neuronal proliferation spurt occurs around the 18th gestational week (Dobbing & Sands 1973) and neurogenesis and the establishment of defined brain structures is essentially complete at the moment of birth (Huttenlocher & Dabholkar 1997). The spurt in neuronal proliferation is closely followed by an increase in glial proliferation around 24th postconceptional week and the whole brain continues to growth until the 3rd or 4th postnatal years (Dobbing & Sands 1973). *Yap1* has been found particularly highly expressed among brain regions with high levels of proliferation (Orr et al 2011) suggesting an important role in driving cell proliferation in this organ. Both proliferation and migration are essential for the establishment of the six layered cortical structure completed just before birth (Rakic 2002). Although synaptic connections start being established before birth, maximal synaptic density is reached between 2 and 16 months of age depending on the cortical region (Huttenlocher & Dabholkar 1997), and myelinisation peaks between 6 months to 2 years, though continues through adulthood (Jiang & Nardelli 2015).

In the present study we analyse transcriptome data for 150 human brain samples corresponding to 15 different brain regions taken from male and female individuals at 10 developmental time points spanning from 12 weeks prenatal development to 3 years of age. By constructing coexpression networks from prenatal and postnatal samples and identifying *yap1* interactions specific to postnatal stages, we aimed to dissect the gene interaction networks for *yap1* in the context of a lower proliferation state which remain poorly understood.

Methods

Human brain gene expression data

RPKM normalized gene expression data was obtained from the Developing Human Brain Atlas (Hawrylycz et al 2012, http://www.brainspan.org/). A group of 150 samples representing 15 different gene structures across 10 developmental ages was selected (Structures: primary auditory cortex (A1C), amygdaloid complex (AMY), cerebellum or cerebellar cortex (CBC), dorsolateral prefrontal cortex (DFC),

hippocampus (HIP), posteroventral (inferior) parietal cortex (IPC), inferolateral temporal cortex (area TEv, area 20) (ITC), primary motor cortex (area M1, area 4; M1C), anterior (rostral) cingulate (medial prefrontal) cortex (MFC), orbital frontal cortex (OFC), primary somatosensory cortex (area S1, areas 3, 1, 2; S1C), posterior (caudal) superior temporal cortex (area 22c) (STC), striatum (STR) and primary visual cortex (striate cortex, area V1/17; V1C); Ages: 12 pcw, 13 pcw, 16 pcw, 21 pcw, 24 pcw, 4 months, 6 months, 1 , 2 and 3 years old). A further quantile normalization was performed and genes with no variance across all samples removed, resulting in a gene population of 21711 genes.

Coexpression networks

Pearson correlation coefficient (r) between the gene expression levels of *yap1* and the expression of every other gene was calculating, and the collection over 21000 coefficients was referred as a *yap1* coexpression profile. Highly coexpressed genes were defined as those with an r >= 0.75. Separate coexpression networks were built for prenatal and postnatal samples.

Similarity between yap1 coexpression profiles

Numeric p value to calculate the significance of the difference in *yap1* prenatal and postnatal coexpression networks was obtained by a simulation. For this we pooled the 75 prenatal and 75 postnatal samples and then randomly divided all 150 samples in two equally sized sets. From each set a *yap1* coexpression profile was constructed and the similarity between the pair of profiles was measured by calculating a Pearson correlation coefficient (r). The same process was repeated a 1000 times, thus obtaining 1000 correlation coefficients that were finally compared to the similarity (r) between prenatal and postnatal *yap1* coexpression profiles. Alternatively, we divided all prenatal samples (n = 75) into five random non-overlapping groups, building a *yap1* coexpression profile from each group. Accordingly five coexpression profiles from non-overlapping groups of postnatal samples were obtained. Next, the similarity between each pair of *yap1* coexpression profiles was calculated through a Pearson correlation. The similarity between coexpression profiles was grouped according to whether they were built from expression data from the same or different developmental stage.

Functional enrichment analyses

Overrepresentation of genes associated with in particular functional groups was assessed with enrichment analysis using WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) (Wang et al 2013, Zhang et al 2005). Enrichment analyses were performed for gene annotations to KEGG pathways, TF binding sites, miRNA targets and GO terms. Parameters used for all enrichment analysis were as follows: reference set of over 21,000 genes with gene expression, hypergeometric test as statistical method, Benjamini Hochberg multiple test adjustment, significance level under 0.05 and 2 as the minimum number of genes for a category.

Yap1 knockdown in human spheroids transcriptome

Spheroids of human cells were grown as previously described by Porazinski et al (Porazinski et al 2015). Deep sequencing was achieved using Illumina HiSeq technology. The overall quality of the reads was evaluated with fastQC, then in order to remove any contamination of adapter sequences 100 nt reads and poor quality reads, a further trimming was accomplished using Trimmomatic (Bolger et al 2014), reads with at least 36 bases where retained for further analysis. Reads were aligned using TopHat2 (Kim et al 2013) with both the complete genome and transcriptome sequences (GTF was obtained from Ensembl). Reads mapping multiple locations were removed from the analysis, the remaining counts were summarized by ensemble gene id using GenomicFeatures and GenomicAlignments packages for R in Bioconductor (Lawrence et al 2013).

Differential expression analysis of yap1 knockdown in human spheroids

Expression data as counts per ensembl gene id was further TMM normalized before comparing gene expression between control shRNA and yap1-shRNA in the human spheroids. Genes were referred as differentially expressed if the Benjamini-Hochberg p value was below 0.01. This resulted in 1316 genes upregulated and 819 downregulated genes at the yap1 knockdown. From these 988 and 669 genes respectively had expression in the human brain.

Overrepresentation of YAP1 known interactions and DE genes

All genes coding for the proteins that interact with YAP1 described by Hauri et al. (Hauri et al 2013) for which they were human brain expression available were selected. Next, the proteins were mapped to their corresponding ensemble gene ids through *uniprot* entry name. Then, the intersection between the genes coexpressed to *yap1* and the genes that physically bind to YAP1 was measured. To assess whether the observed intersection was bigger than expected by chance, the intersection between a random samples of genes (same size as the number of coexpressed genes) and YAP1 interactors was measured, repeating the process 10000 times, and thus allowing us to calculate a numeric p value. *Yap1* coexpressed genes were treated separately, depending on whether the genes were coexpressed to *yap1* exclusively during prenatal or postnatal period or in any developmental stage. A similar analysis was done measuring the intersection between the genes coexpressed to *yap1* in the human brain and the genes differentially expressed in the human spheroids where *yap1* was interfered by a shRNA compared to the control. Differentially expressed genes were treated separately depending on whether they were up or down regulated in the yap1-shRNA human spheroids.

Results

In order to investigate the genomic interactions of *yap1* in a non-proliferative state we constructed coexpression networks for the *yap1* gene from RNA-seq human brain data available at the Developing Human Brain Atlas (Hawrylycz et al 2012, http://www.brainspan.org/), 15 different structures across 10 ages spanning from postconceptional week 12 to 3 years were further divided into 2 subsets, according to whether they correspond to prenatal or postnatal periods. For each of these subsets we measured the coexpression level of *yap1* with every other single gene through a Pearson correlation coefficient. Then, those genes displaying a high degree of coexpression with *yap1* were considered as *yap1* interactors ($r >= 0.75$). We found that a large proportion of these genes are exclusively highly coexpressed with *yap1* either at prenatal or postnatal stage (figure 1).

In order to assess if the observed differences in the coexpression networks in prenatal and postnatal stages for the *yap1* gene are larger than expected by chance, the similarity in *yap1* coexpression in prenatal and postnatal stages was compared to sets of samples divided randomly. For this, each time two *yap1* coexpression profiles were constructed, each of them based on a random selection of half of the samples, then a linear correlation between the profiles was calculated. The same process was repeated a thousand times. The resulting distribution was then compared to the similarity of coexpression profiles in prenatal and postnatal states ($r = 0.3834715$). We found that, as shown in figure 2a, the difference between *yap1* coexpression profiles is significantly larger when the samples are initially divided in pre and postnatal than expected from our random simulation ($p < 0.001$).

As a complementary approach the expression data was divided into 10 smaller non-overlapping groups of 15 samples from either prenatal or postnatal stage. Based on each group of samples a *yap1* coexpression profile was calculated, thus obtaining 10 different profiles. Next, the similarity (r) between each pair of profiles was obtained and based on the distance (1 - r) and a dendrogram was constructed. If the differences observed between *yap1* coexpressed genes at prenatal and postnatal stages were a spurious result we would expected a random clustering of the *yap1* coexpression profiles build from prenatal and postnatal. However, we observed a main division on the dendrogram between the *yap1* coexpression profiles according to their developmental stage (figure 2b). As summarized in figure 2c, we found that similarity of *yap1* coexpression profiles is higher when *yap1* coexpression profiles are constructed based on non-overlapping samples within the same stage than between stages (within prenatal profiles vs. between prenatal and postnatal profiles $t = 8.36$, $p = 5.85 \times 10^{-7}$; within postnatal profiles vs. between prenatal and postnatal profiles $t = 7.73$, $p = 7.62 \times 10^{-6}$).

So far, we demonstrated that *yap1* is associated by coexpression to different gene sets at particular periods of brain development and that the sharp division between these gene populations before and after birth is not expected by chance. Proteins that work together tend to be localized at the same cellular compartments. If the populations of *yap1* coexpressed genes were to be participating in the same function(s) we would expect that the proteins coded in these genes be preferentially localized in defined cell

parts. As an approach of measuring co-localization, we performed an enrichment analysis in gene ontology annotations using the Webgestalt toolkit. A subset of these gene ontology annotations refers to the cell component where the gene products are localized. Remarkably, we not only found a large number of categories overrepresented among the *yap1* coexpressed genes, but also little overlap between the compartments significantly enriched between developmental stages, being the cytoplasm the only compartment significantly enriched in both populations (figure 3). While *yap1* coexpressed genes during prenatal period tend to localize more than expected within the organelle lumen and spatially close to chromosomes, genes coexpressed at postnatal stage are markedly overrepresented at the cell periphery, more precisely at the cell surface, focal adhesion and cell-substrate adherens junction. These results support that yap1 coexpressed genes at different developmental periods in the human brain are involved in different functions.

To further examine these *yap1* coexpressed gene populations, we focused our attention in the biological process terms within the gene ontology enrichment analysis. For each set of temporally specific *yap1* coexpressed genes we found 40 specific biological processes overrepresented (figure 4a). Interestingly, a completely different set of biological process were found enriched among the genes coexpressed to *yap1* prenatally and those coexpressed with *yap1* at a later stage. At prenatal stage *yap1* associates are mainly overrepresented in cell cycle related functions, particularly in its regulation and progression. More specific process such as cell and nuclear division, chromosome condensation and segregation as well as DNA replication were also found being significantly enriched (figure 4a, left panel). In contrast, functions related to development, cell migration, motility, differentiation, signalling, and communication were found significantly enriched among the genes coexpressed to *yap1* exclusively during postnatal development (figure 4a, right panel).

As an independent additional functional classification we used the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways to perform an enrichment analysis. As shown in figure 4b, consistent with our previous result, cell cycle pathway is the most significantly overrepresented among *yap1* prenatal specific coexpressed genes. On the other hand, *yap1* coexpressed genes specific of postnatal period appeared to be enriched predominantly in focal adhesion and in the regulation of the

actin cytoskeleton, followed by other pathways related to metabolism, endocytosis, ECM-receptor interaction and TGF-beta signalling. Even though both gene populations are overrepresented in pathways in cancer, adherens junction and fatty acid metabolism, the majority of pathways enriched among *yap1* coexpressed genes are stage specific.

Genes that participate in the same function are likely to be regulated in the same way. Both transcription factors and microRNAs are the most ubiquitously regulatory factors in multicellular genomes (Hobert 2008). We carried out an enrichment analysis to investigate potential shared regulation of *yap1* coexpressed genes. A transcription factor binding site for an unknown protein (AACTTT) together with binding sites for SP1 and E2F were found beyond expectations among the genes exclusively coexpressed to *yap1* during prenatal period. Motifs were also found significantly overrepresented within the genes uniquely coexpressed to *yap1* at the postnatal period, among those for known transcription factors were FREAC2 (FOXF2), ATF4, OCT1, HFH8, PAX4 and HOXA3, plus another 5 motifs that bind unknown transcription factors (figure 5a). Additionally, a significant excess of miRNA target genes were found among the genes exclusively coexpressed to yap1 after birth, being the most overrepresented sequences recognized by MIR-124A, MIR-200B, MIR-200C, MIR-429, MIR-17-5P, MIR-20A, MIR-20B, MIR-106A, MIR-106B and MIR-519, followed by tens of other miRNAs (figure 5b).

Ours results clearly show that *yap1* is not only associated to a particular sets of genes at specific developmental points, but also that this genes are preferentially localized in cell components and engaged in non-overlapping functions in the prenatal and postnatal developmental periods. Furthermore, there is a strong indication that *yap1* coexpressed genes share a common regulation within the same developmental stage, as shown by the significant enrichment of targets of specific transcription factors and microRNAs. There is also a clear distinction on the potential regulators of the genes coexpressed to *yap1* before and after birth. Taken together, these results strongly suggest the involvement of *yap1* in different functions depending on developmental stages of the human brain.

There is not a complete correlation between mRNA expression levels and protein abundance, either due to relevant biological or technical reasons (Maier et al 2009). Nevertheless, an overlap between the genes coexpressed to *yap1* and proteins previously found to physically interact with the YAP1 protein would provide supporting evidence for the interactions we have identified. For this, all genes coding for proteins that interact with YAP1 were obtained from a recent work from Hauri et al. (Hauri et al 2013) were they described the interaction proteome of human Hippo signalling pathway, among which YAP1 is a known downstream effector. Then, separately for the genes coexpressed to *yap1* at each stage, the intersection with the genes coding for the proteins that interact with YAP1 was measured. A total of 15 genes were found to overlap between the genes coexpressed to *yap1* and the YAP1 protein interactors. To assess whether the observed overlap was higher than expected by chance, we estimated the size of the intersection by taking 10000 random samples of genes of the same size as the number of genes coexpressed to *yap1* and then for each sample we measured its overlap with genes coding for YAP1 interactors. Based on the distribution of the resulting values we estimated a numeric p value. A significant overlap was found between the genes coexpressed to *yap1* and YAP1 protein interactors at prenatal period ($p = 0.0013$) and after birth ($p = 0.0666$, shown in Table 1).

The fact that two genes are coexpressed does not imply a directional regulation of one gene to the other, it merely reflects the synchrony in increases and reductions of expression levels, thus unlike protein-protein interactions provides only an indirect association between genes. However, if among *yap1* coexpressed genes there were some downstream effectors of *yap1* we would expect that some of them, at least more than expected by chance, would to be differentially expressed when *yap1* is knockdown. In this case we asked if among *yap1* coexpressed genes we find an overrepresentation of differentially expressed genes in a YAP1 knockdown.

To test this idea, *yap1* was interfered by a small hairpin RNA at human cell spheroids. RNA was extracted and sequenced for 3 biological replicates of cells transformed with a yap1-shRNA and another 3 for the corresponding control shRNA. Based on the RNA-seq expression data we were able to perform a differential expression analysis using a linear model approach (Ritchie et al 2015). This method allowed us to identify

1316 up regulated and 988 down regulated genes at the *yap1* knockdown. From which only 819 and 669 (up and down regulated respectively) had detectable expression in the human brain. The overlap between differentially expressed (DE) genes in the *yap1*-shRNA and the *yap1* coexpressed genes was measured separately for up and down regulated genes, and for genes highly coexpressed to *yap1* at specific stages. To determine how significant the observed overlap were, an estimated overlap and corresponding numeric p value was calculated based on the overlap from 10,000 random samples of equal size as the *yap1* coexpressed genes population. Remarkably, an overrepresentation of DE genes in every single subpopulation of *yap1* coexpressed genes was detected, with no particular preference for up or down regulated genes (Figure 6, Table 2).

Discussion

Here we have used a coexpression network approach to identify gene populations that are associated to *yap1* in a context dependent manner, contrasting two brain developmental stages marked by different degrees of cell proliferation thus allowing us to gain insights on the architecture of the gene network of *yap1* in non-proliferative functions which have remained poorly understood.

To start with, we identified the gene populations of genes highly coexpressed to *yap1* particularly at prenatal and postnatal stages of brain development. The gene populations associated to *yap1* at prenatal and postnatal stage had little overlap (Jaccard index = 0.06, r >= 0.75, figure 1). When compared to a random partition of the samples there is no other comparison where this difference is particularly marked as when developmental periods before and after birth are compared, this was supported by a two different and complementary simulation approaches (figure 2). This difference in *yap1* coexpression profiles between prenatal and postnatal development is likely to reflect change in the relevance of *yap1* interactions in the context of proliferative and non-proliferative developmental stages in the brain. As a further support of our coexpression network approach, we were able to identify significantly overrepresented proteins that directly bind to YAP1 among the prenatally *yap1* coexpressed genes, and an almost significant enrichment within the corresponding postnatal gene population (Table1). It is reasonable that the difference between the

results between prenatal and postnatal *yap1* coexpressed genes is due to the fact that HEK cells that Hauri et al. used to determine YAP1 protein-protein interactions are highly replicative, thus, biased towards proteins present while YAP1 is likely to be involved in a proliferative role.

By performing an enrichment analysis of cellular components, the only common significantly enriched term was the cytoplasm, a rather general and large cellular compartment, while more specific, smaller and importantly non-overlapping components were overrepresented only at precise developmental periods, such as chromosome part and replication fork at prenatal development, and cell adhesion and cytoplasm-membrane bounded vesicle at the postnatal stage. Functional enrichment analyses of gene Ontology biological processes and KEGG pathways confirmed our expectations, prenatally *yap1* coexpressed genes tend to participate in a more proliferative related functions (i.e. cell cycle), opposed to a postnatal development where this functional signature is no longer detected, and instead a completely different set of functions is revealed. As an example, exclusively coexpressed to *yap1* during prenatal brain development were the proliferating cell nuclear antigen (PCNA), along with two cyclins (CCNA2 and CCNB1), a cyclin dependent kinase (CDK2), cell cycle division proteins (CDC14A, CDC20, CDCA5) among over forty genes directly associated to cell cycle.

The apparently heterogeneous functions enriched among *yap1* coexpressed genes at postnatal stage are actually coherent and include elements mediating the interactions between cells (gap junction, cell junction organization, cell communication and regulation of cell communication), and the extracellular matrix (focal adhesion, ECM receptor interaction), denoting the role of postnatal *yap1* coexpressed genes as both sensors of mechanical force and external signals as well as components integrating those signals within the cell (cell surface receptor signalling pathway, enzyme linked receptor protein signalling pathway, signalling, regulation of signalling, single organism signalling, TGF beta signalling pathway) which might derive in cellular responses including developmental processes, migration, differentiation, morphogenesis and even changes in some metabolic processes (organic acid biosynthetic process, fatty acid metabolic process, amino acid metabolism).

During normal morphogenesis cells must interact with others and their surroundings. It has been proposed that during development different cells can be spatially sorted according to their surface tensions (Foty et al 1996). Accordingly, cell types characterized by a greater cohesion are sorted within cell with minor cohesion reaching an equilibrium state once arranged (Foty et al 1996). Differences in tension and cell cohesion between cell types depend on a combination of adhesion between cells, between cell and the extracellular matrix as well as the response of the internal architecture to external forces in order to maintain a specific shape (Foty & Steinberg 2005, Lecuit & Lenne 2007). In the present work, coexpressed to *yap1* during postnatal brain development we found several genes coding for the proteins that mediate cell-cell contacts, including to ocludin (OCLN) and tricellulin (MARVELD2), the first allows bicellular tight junctions while latter as its name indicates allows the contact between three cells and have a critical in the formation of the epithelial barrier (Ikenouchi et al 2005). For instance tissue surface tension increase as the expression levels of cadherins gets higher (Foty & Steinberg 2005), here we have found one of these calcium-dependent cell adhesion proteins *cdh20* and the genes coding for the cadherin-associated protein (CTNNA1) highly coexpressed to *yap1* during postnatal development. YAP1 itself has been propose to regulate tissue tension upstream ARHGAP18 by altering the actin and fibronectin assembly in human spheroids (Porazinski et al 2015), however the role of other GTPase-activating proteins has not been excluded. In accordance to these findings, we found *arhgap31* to be highly coexpressed with *yap1* in the postnatal brain; this gene codes for another GTPase-activating protein which mutations has been linked to Adams-Oliver syndrome, a rare disease involving congenital scalp defects and terminal transverse limb defects (Isrie et al 2014). We also found genes coding for fibronectin (*fn1*) and moesin (*msn*) to be highly co-expressed with *yap1* at postnatal stage. While FN1 is a major glycoprotein of the extracellular matrix that typically binds to integrin molecules (Busk et al 1992), MSN is a notable linker between the plasma membrane and actin cytoskeleton at the cell cortex (Amieva & Furthmayr 1995). Both *fn1* and *msn* (together with *arhgap19*) has been found down regulated by the microRNA miR-200c (Howe et al 2011), whose binding sequence was found overrepresented in yap1 coexpressed genes (Figure 5b, right panel). Some members of the integrin gene family *itga1*, *itga6*, *itgb1* and *itgb5*, along with erzin (*erz*) which product forms a protein complex with that of *msn*, and vimentin (*vim*) a member of the intermediate filament, are also part of the machinery

that permit a cell to respond to its environment (Gary & Bretscher 1993).These genes were all found to be highly coexpressed to *yap1* in a postnatal non-proliferative context. Notably, *erz* and *vim* were found differentially expressed at human spheroids when *yap1* was knockdown.

Beyond the actual sets of functions identified for each set of *yap1* coexpressed genes, we further investigated the potential common regulators of these gene groups with both an enrichment of transcription factor binding sites and mi-RNA binding sites. These analyses revealed mostly different TF binding sites and in the case of mi-RNA, the overrepresentation was only among the *yap1* coexpressed genes at postnatal stage, signifying a predominant regulatory role of mi-RNAs over *yap1* associates in a non-proliferative context.

As a co-activator, YAP1 interacts with several transcription factors to promote or inhibit the expression of a large number of target genes. Is worth mentioning that based on the brain expression data analysed here, we found *tead1* and *tead2* coexpressed to *yap1* exclusively during prenatal development, suggesting that particularly at this developmental stage they may be working together to promote the transcription of other genes. However we cannot rule out the possibility the involvement of TEAD proteins in other non-proliferative functions.

In summary, by applying a coexpression network approach to analyse the role and the interactors of the *yap1* gene in highly proliferative and low proliferative prenatal and postnatal brain development stages we have identified a distinct set of gene interactors to each stage. The functional analyses of these sets of genes have revealed a marked difference in the functions of the genes related to *yap1* at each stage. Our findings reveal that interactors for *yap1* in both stages are enriched in different cell compartments and identify novel molecular interactors of *yap1* which might be key to *yap1*'s roles in non-proliferative functions.

Conclusions

Based in a differential coexpression approach involving human brain we were able to identify a significant change in the way *yap1* is coexpressed to other genes before and

after birth. Being YAP1 a major hub in signal transduction, the contrast between the prenatal and postnatal associates in terms of their localization, function, and potential regulation (by transcription factors and microRNAs) allow us to separate the proliferative and non-proliferative role of *yap1* in the human brain. Furthermore, known protein-protein interactors of YAP1 and down regulated genes when *yap1* was knockdown in human spheroids. Strongly suggesting that at this latter stage *yap1* and its associates is involved in integrating external cues, and conceivably involved in the maintenance of tissue tension.

Tables

Table 1. Overlap between *yap1* coexpressed genes and YAP1 known protein interactors (Hauri et al 2013). We obtained different gene subpopulations of genes coexpressed to *yap1*, depending if the genes were highly coexpressed to *yap1* either only during prenatal or postnatal development or at both developmental stages. In the other hand, we were able to match the human brain expression data available with 225 genes coding for known physical YAP1 interactors. For each subpopulation of *yap1* coexpressed genes we test if there was a significant enrichment of genes coding for YAP1 protein interactors based on a simulation analysis (*see Methods*).

Yap1 coexpressed genes (r >= 0.75)	Number of genes	Observed YAP1 protein interactors	Expected	SEM	Numeric *p* value
At any stage	494	15	5.1247	2.2272	**0.0001**
Only during prenatal stage	180	8	1.8748	1.3692	**0.0013**
Only during postnatal stage	282	6	2.8797	1.6636	0.0666
At both prenatal and postnatal stage	32	1	0.3372	0.5743	0.2886

Table 2. Overlap between *yap1* coexpressed genes and differentially expressed (DE) genes in human spheroid cells *yap1* knockdown. Gene expression was compared between cells where *yap1* expression was repressed through the expression of a yap1-shRNA and cells where a control shRNA was introduced. Gene differential expression was tested using *limma*, as a result we found a total of 1316 and 819 genes up and down regulated respectively in the yap1-shRNA cells (*see Methods*). For each subpopulation of *yap1* coexpressed genes (genes coexpressed at a particular developmental stage), we test if there was an enrichment of genes differentially expressed in human cell spheroids *yap1* knockdown. Enrichment was tested separately for up and down regulated genes:

Yap1 coexpressed genes (r >= 0.75)	Number of genes	UP regulated	Expected	SEM	Numeric *p* value
At any stage	494	54	29.1365	5.1689	**0.0001**
Only during prenatal stage	180	20	10.8188	3.1316	**0.0059**
Only during postnatal stage	282	27	16.4665	3.8615	**0.0084**
At both prenatal and postnatal stage	32	7	1.8841	1.3194	**0.0021**

Yap1 coexpressed genes (r >= 0.75)	Number of genes	DOWN regulated	Expected	SEM	Numeric *p* value
At any stage	494	49	19.7786	4.3146	**0.0001**
Only during prenatal stage	180	23	7.3648	2.6219	**0.0001**
Only during postnatal stage	282	22	11.1164	3.2342	**0.0020**
At both prenatal and postnatal stage	32	4	1.289	1.0978	**0.0354**

Figures

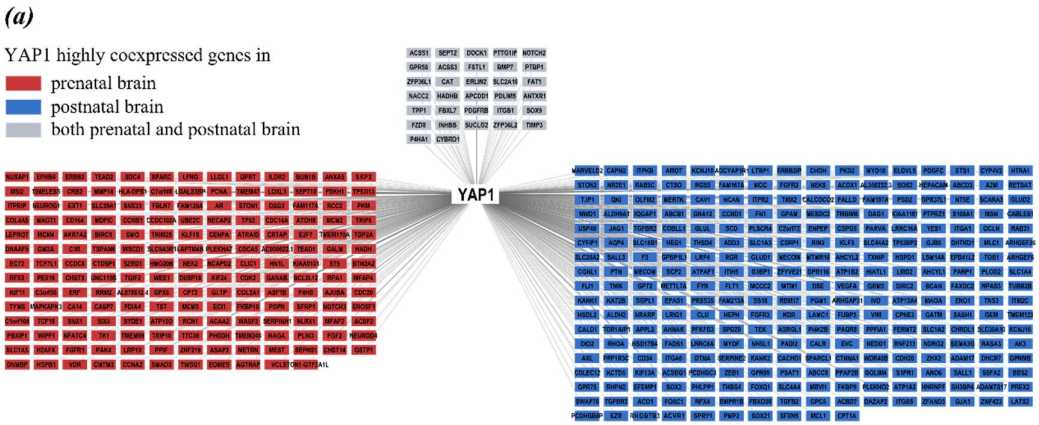


Figure 1. *Yap1* coexpressed genes at different developmental stages in the human brain. ***(a)*** The nodes in the graph denote genes and the edges their coexpression with *yap1* (r >= 0.75). Colour code indicate whether the high level of coexpression in maintain only in prenatal (pink), postnatal (blue) or at both developmental stages (grey). *Yap1* is coexpressed with an almost different set of genes at each stage.

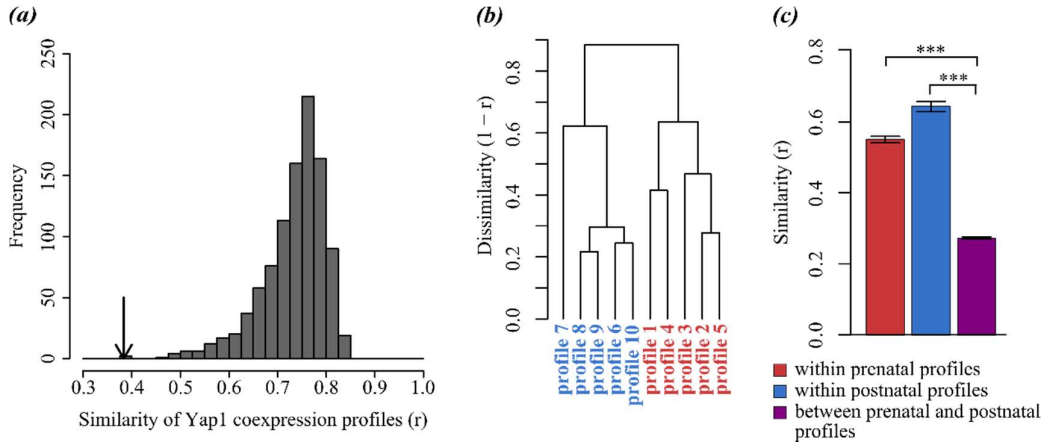


Figure 2. The observed differences in *yap1* coexpression profiles between developmental stages are higher than expected by chance. ***(a)*** Histogram showing the resulting 1000 correlations measuring the similarity between each pair of *yap1* coexpression profiles constructed from a random division of all samples, while the arrow indicates the significantly reduced similarity between the prenatal and postnatal *yap1* coexpression profiles ($p < 0.001$). In a complementary approach, all prenatal samples were randomly divided in 5 non overlapping groups of 15 samples, based on each group the Pearson correlation between the expression of every single gene and the expression of *yap1* was obtained. Thus obtaining five coexpression profiles for *yap1* (profile 1 to 5) based on prenatal samples, likewise we built five coexpression profiles for *yap1* based on all postnatal samples (profile 6 to 10). Lastly, the similarity between the resulting *yap1* coexpression profiles was obtain as a Person correlation. ***(b)*** Dendrogram shows the dissimilarity (1 - r) between *yap1* coexpression profiles built from either prenatal (salmon) or postnatal (blue) non-overlapping samples. ***(c)*** Barplot showing the average similarity (r) within *yap1* coexpression profiles built from prenatal samples (salmon), within *yap1* coexpression profiles built from postnatal samples (blue) or between *yap1* coexpression profiles built from samples of different stages (purple), error bars denote the standard error of the similarity between pairs of profiles (within prenatal profiles vs. between prenatal and postnatal profiles $t = 8.36$, $p = 5.85 \times 10^{-7}$; within postnatal profiles vs. between prenatal and postnatal profiles $t = 7.73$, $p = 7.62 \times 10^{-6}$).

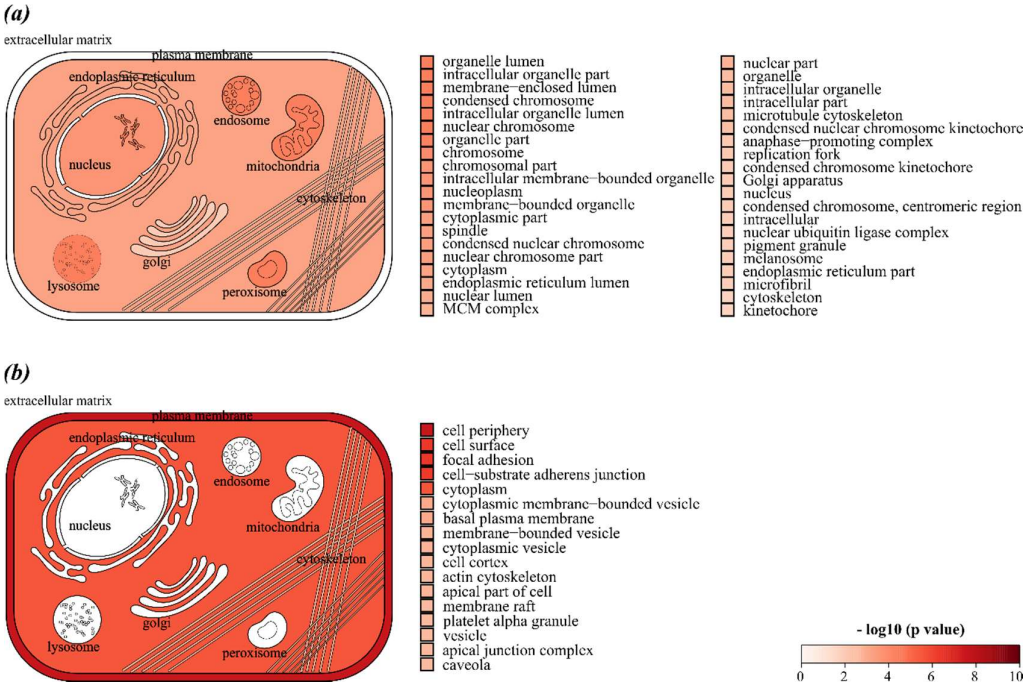


Figure 3. The products of *yap1* coexpressed genes at different stages localize preferentially at specific cell components. Using the online WebGestalt toolkit we performed a Gene Ontology enrichment analysis for all the genes that are highly coexpressed ($r >= 0.75$) with *yap1*. The cell diagrams summarize the significantly overrepresented cellular component GO terms after Benjamini Hochberg correction ($p_{BH\ adjusted} < 0.05$) among ***(a)*** *yap1* highly coexpressed genes only during prenatal development and ***(b)*** *yap1* coexpressed genes only during postnatal stage. Squares next to each gene ontology indicate the minus logarithm of the adjusted p value, as the red intensity increases the more significant overrepresentation of genes among that particular GO term.

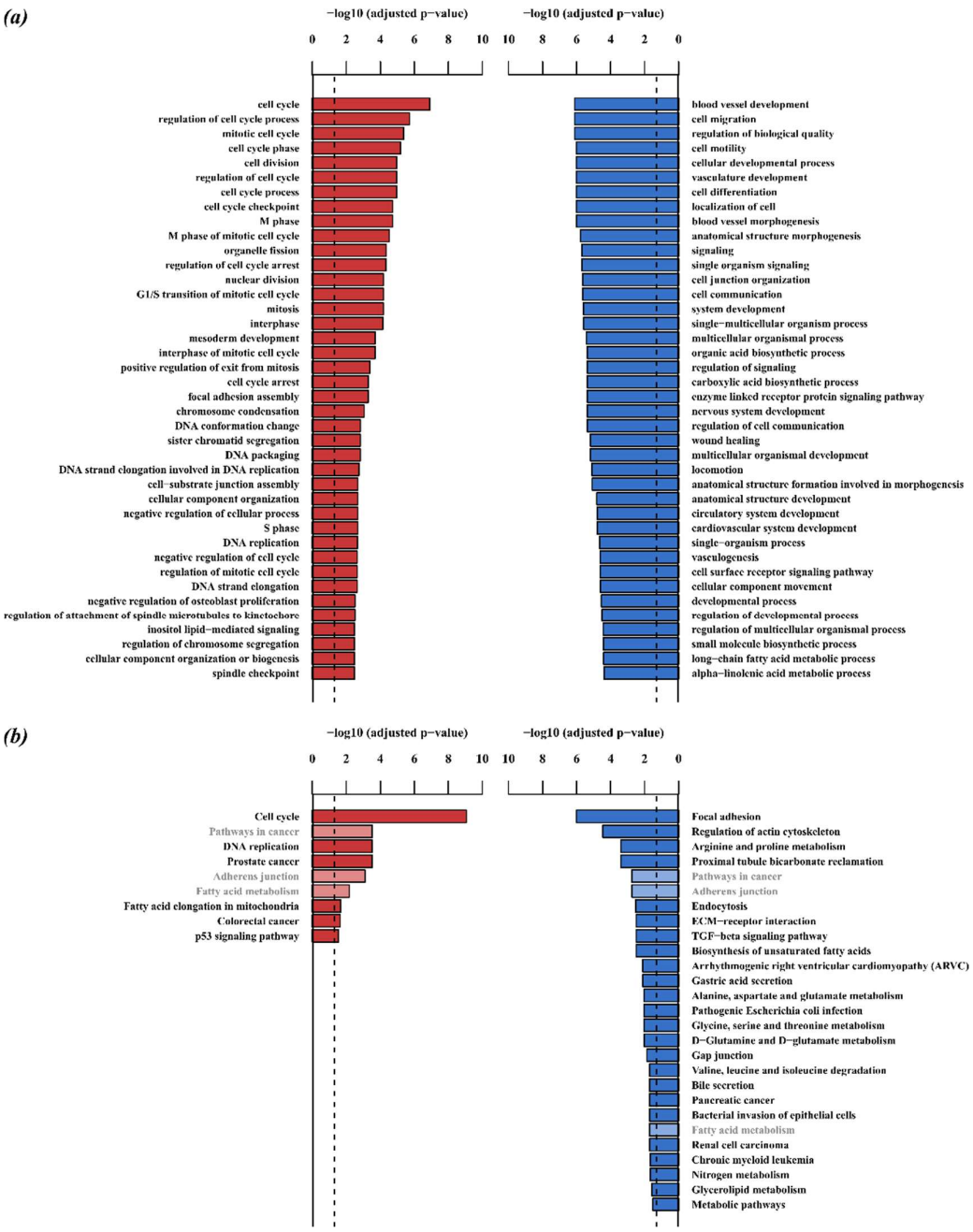


Figure 4. GO and KEGG enrichment analysis reveals clear functional differences among *yap1* coexpressed genes at different developmental stages. Barplots represent the minus logarithm of Benjamini Hochberg adjusted p value for categories significantly enriched among genes highly coexpressed with *yap1* only during prenatal (left) or postnatal period (right). Dashed line mark the threshold above which categories were deemed significant ($p_{\text{BH adjusted}} < 0.05$). ***(a)*** Shows biological process GO terms significantly enriched while ***(b)*** denote overrepresentation of genes among KEGG pathways. Black labels denote GO terms and KEGG pathways exclusively enriched at a particular developmental stage.

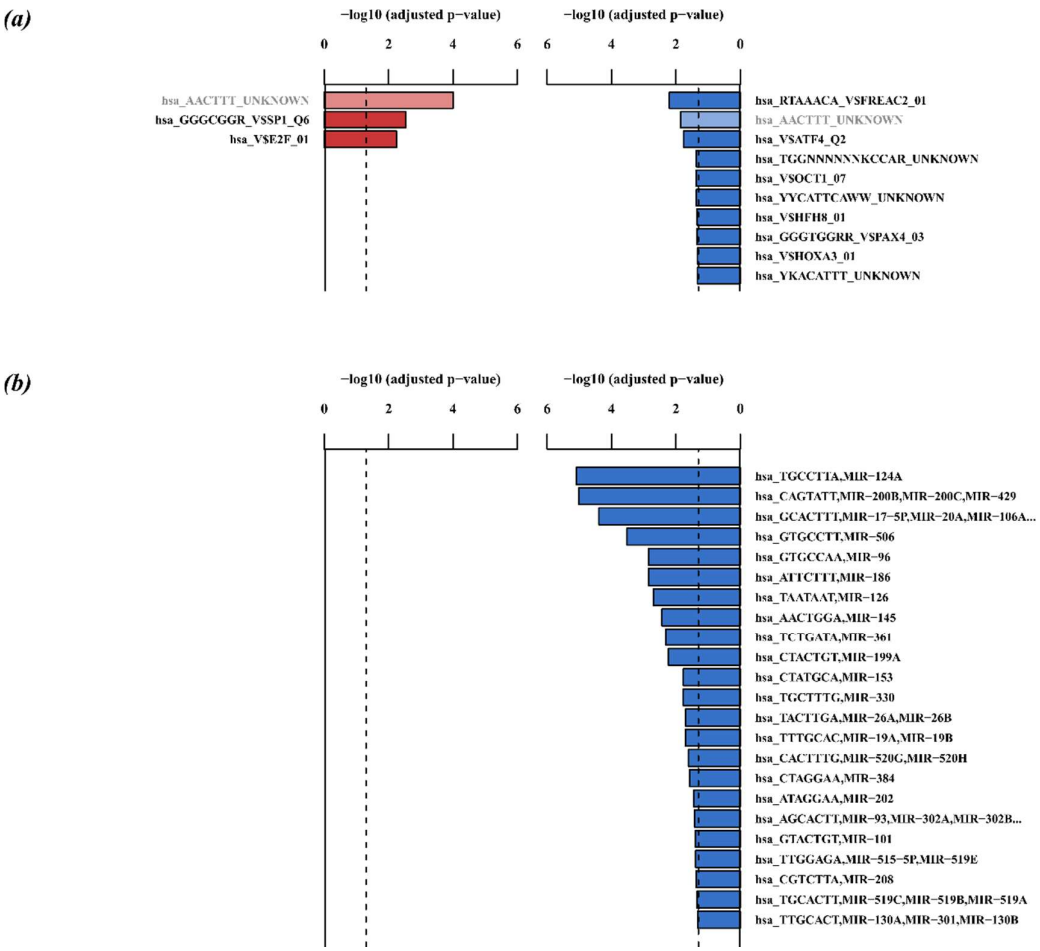


Figure 5. TF and miRNA targets enrichment analysis of *yap1* coexpressed genes at different developmental stages shed light into a differential regulation. Barplot represent the minus logarithm of Benjamini Hochberg adjusted p value for categories significantly enriched among genes highly coexpressed with *yap1* only during prenatal or postnatal period (left and right, respectively). Dashed line mark the threshold above which categories were deemed significant ($p_{BH\ adjusted} < 0.05$). ***(a)*** Barplot shows significantly overrepresented transcription factor binding motifs while ***(b)*** denote enrichment of target genes containing a sequence recognized by a specific miRNA. Black labels denote TF and miRNA binding sites exclusively overrepresented at a particular developmental stage.

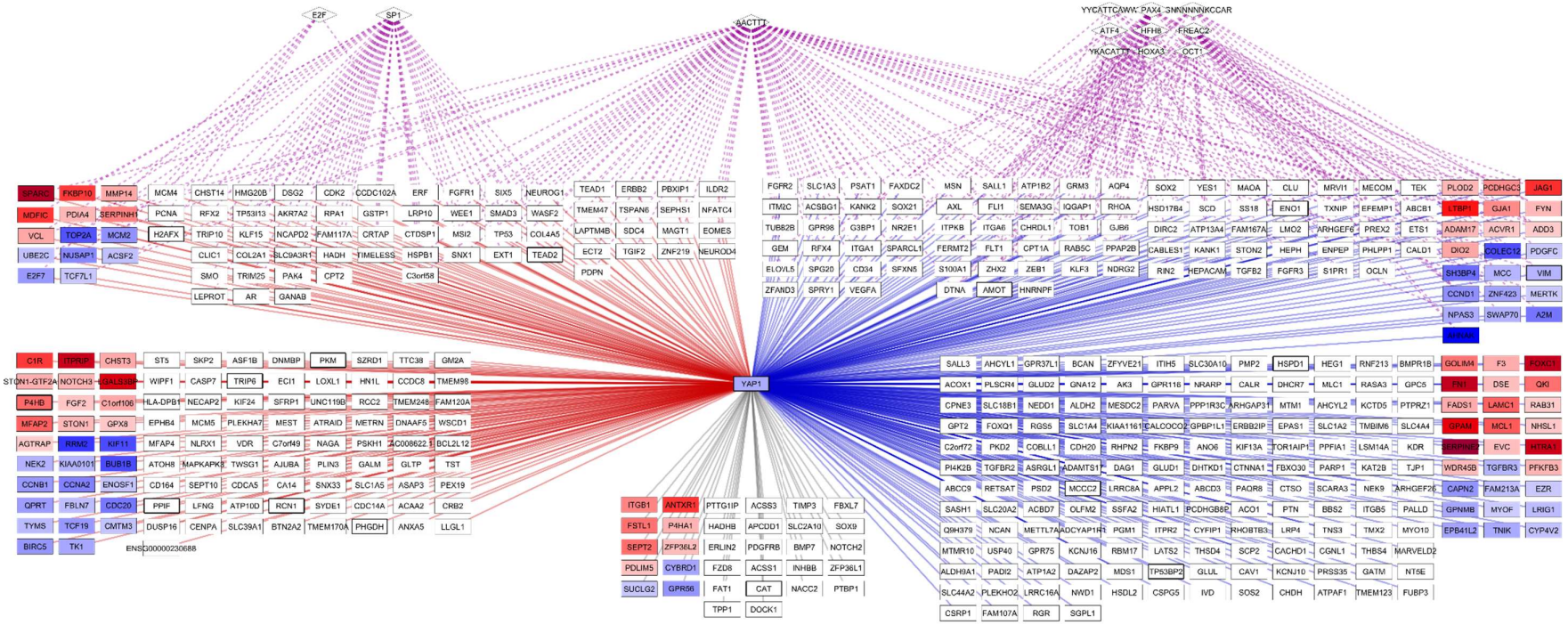

E2F
SP1
AACTTT
YYCATTCAWW_PAX4_SNNNNNNKCCAR
ATF4
HFH8
FREAC2
YKACATTY
HOXA3
OCT1
YAP1
gene highly coexpressed to yap1 (r > 0.75)
up regulated gene in yap1 KD
down regulated gene in yap1 KD
gene product physically interact with YAP1
TF associated to a binding site significantly overrepresented
high coexpression with yap1 only during prenatal development
high coexpression with yap1 only during postnatal development
high coexpression with yap1 across development
potential regulation of expression by transcription factor

Figure 6. *Yap1* coexpression network integration. ***(a)*** The network represents as rectangles all those genes that are highly coexpressed to *yap1* at different developmental processes. Rectangles are coloured in either red or blue depending on whether they were significantly over or under expressed at the *yap1* knockout in human spheroids. Thicker border line denote genes coding for proteins that are known to bind YAP1 according to Hauri (Hauri et al 2013). Diamonds represent TF binding sites overrepresented among the different populations of coexpressed genes. Solid lines denote a high association between the expression levels of two genes at either prenatal (red), postnatal (blue) or both stages (grey). Purple dashed lines link a gene to its potential regulator given the overrepresented TF binding site (or the overrepresented if there is not any known factors that bind that sequence).

Chapter 4.

The evolution of longer lifespan is associated with size variations in gene families related to immune system function

Jimena Monzón-Sandoval[1, 2], Atahualpa Castillo-Morales[1, 2], Araxi O. Urrutia[1], Humberto Gutierrez[2]

[1] Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

[2] School of Life Sciences, University of Lincoln, Lincoln LN6 7DL, UK

JMS: jms52@bath.ac.uk

ACM: acm39@bath.ac.uk

AOU: a.urrutia@bath.ac.uk

HG: hgutierrez@lincoln.ac.uk

Abstract

Lifespan is a highly variable trait among mammalian species, differences can be counted in orders of magnitude. Evolution of a longer lifespan may be linked to genomic adaptations that reduce damage and/or promote maintenance, increase resistance to stressors, pathogens and/or environmental threads. Still, the nature of genomic changes accounting for such variations in species lifespan are only beginning to be explored (Li & de Magalhaes 2013). Although overall gene number has remained relatively constant over metazoan evolution, whole genome comparisons have revealed dramatic changes in size of individual gene families (Demuth et al 2006, Fortna et al 2004, Hahn et al 2007a, Hahn et al 2007b, Hughes & Friedman 2004, Rubin et al 2000). These variations have been proposed to play a major role in shaping morphological and physiological differences between species. Here we found, a significant enrichment of gene family size (GFS) variations in line with increased lifespan. Using a comparative approach including 28 sequenced mammalian species and 11943 gene families we show parallel changes in GFS and maximum lifespan that are not secondary to known correlates of lifespan (degree of encephalization and neocortex to brain ratio) and are not completely explained by phylogenetic effects. Immune and defence response related functions are overrepresented among these families, and most remarkably gene variants previously associated to longevity in humans. We propose that during evolution of longer lifespan in mammals, underlying genetic adaptations of longevity and defence response mechanisms were in part brought about by changes in the number of gene copies within selected gene families.

Introduction

Lifespan is in part limited by aging, a complex process of accumulation of molecular, cellular, and organ damage, leading to loss of function and increased vulnerability to disease and death. The most accepted theories of ageing fall into two camps, programmed and damage or error theories. Program theories imply that lifespan is limited, presumably an unavoidable by-product of the intrinsic program controlling development and growth (Walker 2011). This developmental program is proposed to depend on changes in gene expression that affect the systems responsible for maintenance, repair and defence responses. Error theories on the other hand emphasize environmental factors and insults compromising the organism's integrity leading to cumulative damage at various levels resulting in aging (Kirkwood et al 2000, Wensink et al 2012).

Comparative studies suggest that variations in maximum lifespan across species reflect intrinsic differences in the molecular machinery governing the ability of organisms to cope with age-related cellular damage and vulnerability to disease (Finch et al 2010, Harper et al 2007, Kirkwood et al 2000, Kourtis & Tavernarakis 2011, Perez et al 2009, Ricklefs 2010, Schumacher et al 2008). Among these studies Ricklefs and Cadena identified a significant genetic contribution to the age of death in captive populations of wild mammals (Ricklefs & Cadena 2008). A general positive relation between the survival of primary fibroblast cultures exposed to various stressors and the corresponding lifespans of the species from which the fibroblast came from has been reported (Finch et al 2010). Furthermore it has been proposed that species specific cellular responses to stress and inflammation may contribute to the evolution of longevity (Finch et al 2010, Iannitti & Palmieri 2011). Comparative studies have been proven to be useful to identify genes and pathways relating accelerated protein evolution and mammalian longevity (Li & de Magalhaes 2013). However other genomic changes accounting for the observed differences in longevity across species remain largely unexplored.

While overall gene number has changed little over the last 800 my of metazoan evolution, recent analyses of gene family evolution have shown instances of dramatic gene family expansion and contraction with frequent gene turnover (Ashburner et al

2000, Demuth et al 2006, Fortna et al 2004, Hahn et al 2007a, Hahn et al 2007b, Hughes & Friedman 2004). The fact that these variations are accentuated in gene families associated with certain functional categories suggests that changes in gene number within gene families may reflect evolutionary responses to specific functional demands (Castillo-Morales et al 2014, Hahn et al 2007a, Hahn et al 2007b, Hughes & Friedman 2004, Kapheim et al 2015). Here we use a comparative genomics strategy covering 28 fully sequenced mammals in order to identify the connections between lifespan and GFS variation.

Methods

Gene family size annotations

Annotated gene families encompassing 28 fully sequenced mammalian genomes were obtained from Ensembl release 76 (Flicek et al 2014). In the context of this annotation, Ensembl families are defined by clustering all Ensembl proteins along with metazoan sequences from UniProtKB. Any given gene family constitutes a group of related genes that include both paralogs within the same species and orthologs and paralogs from other species. GFS represents the total number of genes per gene family. In order to maximize the number of families covered in this study we included all gene families with members present in no less than six of the 42 sequenced mammalian species and excluded any family with no variance in GFS across the subset of 28 species analysed (n = 11943).

Maximum lifespan, encephalization index and neocortex to brain ratio

Maximum lifespan (MLSP) recorded for each species was obtained from the animal ageing and longevity database (Tacutu et al 2013). Encephalization index values (Ei), which measures the brain mass controlling for the allometric effect of body size, were computed as log [brain mass/body mass]. The slope (b) was estimated as 0.64 by Gonzalez-Lagos et al. (Gonzalez-Lagos et al 2010) based on a log–log least squares linear regression of brain mass against body mass over 493 mammalian. Neocortex to brain ratio (Nr) was defined as the ratio of neocortex volume to the volume of the rest of the brain (Beatty & Laughlin 2006, Bush & Allman 2004, Ebinger 1974, Georg

Baron 1996, Hakeem et al 2005, Kruska & Rohrs 1974, Kruska 2014, Pirlot 1981, Pirlot & Jiao 1985, Reep et al 2007, Stephan et al 1981). All MLSP, Ei and Nr values used in this study are presented in Table 1.

Longevity related genes

We used publicly available databases containing genes that are either associated to longevity or that may regulate or influence longevity. Among these, LongevityMap (Budovsky et al 2013) is a collection of gene variants associated to longevity in human populations. GeneDR is a database of genes related to dietary restriction, which has been studied as a conserved mechanism to extend lifespan in multiple organisms (Wuttke et al 2012). REPAIRtoire contain genes involved in DNA repair (Milanowska et al 2011) while DeathBase is a compilation of genes related to apoptosis (Diez et al 2010).

Correlation coefficients of gene family size and phenotypes

Pearson correlations between GFS and MLSP were calculated using R software. In order to determine whether the resulting correlation coefficient distribution was expected by chance, 10000 Monte-Carlo randomizations we performed, the resulting correlation coefficient distribution was averaged, and then compared to the true observed distribution of correlation coefficients for each variable. In order to control the potential contribution of Nr and Ei on the relationship between MLSP and GFS, partial correlation coefficients were computed for each gene family including Nr and Ei as covariates. The resulting partial coefficient represents the contribution of MLSP to the variance in GFS which is not explained by variations in the other two confounding phenotypes.

Phylogenetically controlled regression

In order to further unravel the contribution of each phenotype to GFS, we accounted for the phylogenetic non-independence of taxa on the relationships of morphological traits, phylogenetic independent contrasts (PIC) analysis were used (Felsenstein 1985, Paradis et al 2004). We first obtained residuals of MLSP, by regressing it against the other two phenotypes in a multivariate lineal model. The same regression was

performed on GFS of each family, again, using the other two as predictors. PIC for both the partialized phenotype and the partialized GFS of each family were computed using the ape package in R. Pearson correlation was then carried out between each of these partialized phenotypes and the partialized size of each family. This correlation coefficient reflects the degree of association between MLSP and GFS when both confounding variables and phylogenetic non-independence are controlled. Ultrametric phylogeny of the 28 analysed mammalian species obtained from TimeTree website (Kumar & Hedges 2011) http://www.timetree.org/, last accessed on 05/01/2015

Gene Ontology enrichment analyses

Gene ontology annotations for each species were obtained from Ensembl's Biomart release 76. In the present study, a particular GO term was associated to a family whenever that term was linked to any of its members in any species. Only terms found to be linked with more than 200 families were examined (n = 116). Gene ontology annotations are mainly separated in three groups, a set that include biological processes, a set for molecular functions and finally a set of cellular component localization. We performed a separate GO enrichment analysis for each group of ontology terms. Gene families not annotated to any GO term of a particular group in any species were grouped into a "not annotated" category. Gene families annotated to any GO term with less than 200 families were assigned to a "small terms" category. Enrichment analysis of these GO terms was carried out by counting the number of families assigned to each GO term within the analysed set of gene families. Statistical significance was numerically assessed by obtaining the expected number of families per GO in 1000 equally sized random samples derived from the overall population of gene families.

Results

Here we investigate if size variations of gene families in 28 fully sequenced mammalian species have occurred in line with increased lifespan. Annotated gene families were obtained from Ensembl (Flicek et al 2014) and the resulting 461845 genes in this dataset were distributed among 11943 gene families with an average of 38.67 genes per family across all 28 genomes and the number of copies per gene family

per species ranging from 0 to 292. Only families represented in at least 6 of the 42 sequenced mammalian species were included in this study. We used maximum longevity (MLSP), obtained from a comprehensive collection available in AnAge database (de Magalhaes & Costa 2009), as this measure, unlike mean lifespan, is determined by the intrinsic species specific characteristics that allows them to survive to certain age.

For each gene family we calculated the Pearson correlation between MLSP and the number of genes present for that family across 28 mammals. The resulting distribution of 11943 correlation coefficients showed a highly significant shift towards positive values when contrasted with Monte Carlo simulations of the expected distribution based on random permutations of GFS values across species ($P << 0.0001$, Figure 1A). Relative to an expected zero-valued average correlation, we found 9055 families with correlation coefficient above zero, representing a shift of 3083 gene families from the negative to the positive tail of the distribution, well above chance expectations ($X^2 = 3184.45$, $p \approx 0$). This result demonstrates a significant surplus of gene families displaying a specific positive association between GFS and increased lifespan. Moreover, within those families 528 were found to have a statistically significant association with MLSP after correcting for multiple testing (Benjamini-Hochberg correction $r_{\text{MLSP GFS}} > 0$, $p_{\text{BH adjusted}} < 0.05$). It is worth mentioning that we only observed two gene families with significantly negative association between GFS and MLSP ($r_{\text{MLSP GFS}} < 0$, $p_{\text{BH adjusted}} < 0.05$).

Other phenotypic traits, such as relative brain size and neocortex to brain ratio has been found strongly correlated with lifespan in mammals (Allen et al 2005, Barrickman et al 2008, Gonzalez-Lagos et al 2010) and thus, any potential associations between MLSP and GFS could be secondary to an underlying association between GFS and either encephalization, neocortex to brain ratio, or both. (In this particular set of 28 species the correlation between MLSP and neocortex ratio is $r = 0.861$, while the correlation between MLSP and encephalization is $r = 0.857$). Accordingly, using GFS as the predicted variable, we carried out multivariate regressions including MLSP, neocortex ratio and encephalization (Castillo-Morales et al 2014, Gonzalez-Lagos et al 2010, Gutierrez et al 2011) as covariates and obtained the corresponding partial correlation coefficients for MLSP each of the 11943 gene families. Within the

resulting partial correlations we also found a significant bias towards positive association between MLSP and GFS ($X^2 = 790.699$, $p \approx 5.676 \times 10^{-74}$) and a set of 183 families where GFS is significantly associated to MLSP once the variance of the confounding phenotypes has been removed.

For a given gene family, a strong association between MLSP and GFS could be the spurious result of existing phylogenetic relationships, as in the absence of any selective forces, closely related species will tend to have both similar maximum longevities and similar GFS (Felsenstein 1985, Freckleton et al 2002, Pagel 1999). In order to determine the extent of phylogenetic contribution to the observed shift in the correlation distribution, we carried out Felsenstein's phylogenetically independent contrasts (PIC). For this, first we calculated the residuals of MLSP by regressing it against Ei and Nr in a multivariate lineal model. Similarly for each gene family we regressed the GFS using Ei and Nr as predictors, obtaining both a partial correlation coefficients for MLSP and GFS per gene family. Then for each pair of partialized variables we calculated a PIC and finally a Pearson correlation between each pair was performed. The resulting distribution of correlation coefficients, were both confounding phenotypes and phylogenetic non-independence were controlled, still shows an overrepresentation of positive correlations ($X^2 = 668.226$, $p = 2.428 \times 10^{-147}$) with a large number of families with a significant association between GFS and MLSP after Benjamini Hochberg multiple testing correction ($n = 1059$).

So far, our results show a specific set of gene families where GFS is significantly associated to MLSP, first using simple Pearson correlations, next partial correlations, and then through phylogenetic contrast analysis (Venn Euler Diagram). In order to further assess the nature of the gene families specifically associated with lifespan-related GFS variations, we performed a functional enrichment analysis in Gene Ontology terms (Ashburner et al 2000).

As shown in figure 2A, among those gene families displaying the strongest association between GFS and MLSP, measured with the simple Pearson correlation ($r > 0$, $p_{BH\ adjusted} < 0.05$, $n = 528$) we found a significant overrepresentation of gene families with no biological process annotated, followed by immune response, chemotaxis, defence response to bacterium, cell-cell signalling, defence response, inflammatory response,

response to virus, negative regulation of endopeptidase activity, neuropeptide signalling pathway and finally epidermis development. Interestingly, after removing the variance of the confounding phenotypes (Nr and Ei) within the gene families showing a significant association between GFS and MLSP ($r_{MLSP, Nr Ei} > 0$, $p_{BH\ adjusted} < 0.05$, $n = 183$) neuropeptide signalling pathway, a gene ontology term undoubtedly related to brain function, is no longer enriched, remaining significantly overrepresented immune response and defence response along with ATP catabolic process and cytokine-mediated signalling pathway (Figure 2B). Moreover, when focusing in the population of gene families where GFS are significantly liked to MLSP once we have controlled the effect of the confounding phenotypes and the phylogenetic relatedness ($R_{PIC\ MLSP, Nr Ei\ vs\ PIC\ GFS, Nr Ei} > 0$, $p_{BH\ adjusted} < 0.05$, $n = 1059$), only immune and defence response persist significantly overrepresented (Figure 2C).

When detecting particular molecular functions among the genes from families showing a significant correlation between GFS and MLSP ($r > 0$, $p_{BH\ adjusted} < 0.05$), we found several GO terms significantly enriched, among which chemokine and cytokine activity can be easily linked to the immune function, and interestingly cytochrome-c oxidase activity, decreases in an age dependent manner in rodents (Petrosillo et al 2013), leading to an increase in ROS production, an augment in oxidative stress and feasibly accelerating ageing (Kim et al 2015), and has been shown that the reduction of this activity reduces lifespan in flies (Sohal et al 1995). However, once we controlled for the effect of neocortex ratio, encephalization and account for the phylogenetic relatedness, those gene families which size is positively associated to lifespan ($R_{PIC\ MLSP, Nr Ei\ vs\ PIC\ GFS, Nr Ei} > 0$, $p_{BH\ adjusted} < 0.05$) are no longer enriched in genes with this activity, but instead signal transducer activity and G-protein coupled receptor activity are significantly overrepresented. Finally, in the case of cellular component enrichment analysis, we found an enrichment of genes localized at the keratin and the intermediate filament, plasma membrane, as part of the integral component of plasma membrane, at the lysosome and its membrane, and at the membrane in general.

We investigated if the changes we observed on GFS in line with lifespan, would affect genes already known to play a significant role in influencing or promoting maximum lifespan in mammals. If this was the case, it would strongly suggest that these changes

in GFS are an evolutionary response to the functional demands required to live a longer life.

To this end we used the Longevity Map database, which contains an extensive compilation of human genetic variants associated with longevity in human populations (Budovsky et al 2013). And because their involvement in regulating or influencing lifespan we also used GeneDR a collection of genes associated to caloric restriction which delays degeneration by ageing and extends lifespan in multiple organisms (Wuttke et al 2012); DeathBase which comprise a collection of genes involved in apoptosis (Diez et al 2010) and REPAIRtoire a database that contains genes involved in DNA repair (Milanowska et al 2011). From each database we ascribed all genes to their corresponding gene families. Then, we look if within the gene families showing a significant association between GFS and MLSP ($R_{\text{PIC MLSP, Nr Ei vs PIC GFS, Nr Ei}} > 0$, $p_{\text{BH adjusted}} < 0.05$) there was an enrichment of families containing genes from the Longevity Map database. A significant enrichment of longevity associated genes among our selected gene families ($p = 0.00063$) was observed (Figure 3A). Though we find some overlap with the genes contained in CRgene, Deathbase and REPAIRtoire, there is no significant enrichment of those genes within the families that we identify changing their GFS along with MLSP.

Taken together, once confounding phenotypes and phylogeny has been controlled for, MLSP-associated changes in GFS are significantly enriched in immune and defence response functions, enriched in signal transducer and G-protein coupled receptor activity, and are overrepresented at the filament, as well as lysosome and plasma membrane. Furthermore, we found an enrichment in gene variants that has been associated to longevity in human populations.

Our results also suggest that during evolution of longevity in mammals, the underlying adaptations of immune and defence response mechanisms were, at least in part, brought about by changes in the number of gene copies within selected gene families.

In summary our findings, show that lifespan-associated GFS variations, most likely represent an evolutionary response to the specific functional demands to live a longer life.

Discussion

An extended lifespan is a defining characteristic of the human species. Among mammals, there are strong variations in lifespan but what at the molecular level accounts for the evolution of this trait remains poorly understood. A step in this direction has already been taken, Li et al. have looked for accelerated protein evolution among long-lived mammalian species, hinting towards repair mechanism and the proteasome-ubiquitin pathway (Li & de Magalhaes 2013). Previous multispecies analyses have revealed marked variations in the size of gene family and it has been proposed that differences in GFS could play a role in shaping phenotypic differences among species. By examining variations in GFS and lifespan in 28 sequenced mammalian species we have demonstrated a significant enrichment of GFS variations in line with increased lifespan in mammals and that this enrichment is independent of phenotypic variables known to correlate with lifespan such as degree of encephalization and neocortex to brain ratio. Demonstrating, with a purely correlative approach, a specific association between a phenotypic trait and number of genes within a particular gene family can be difficult as spurious coincidences cannot be ruled out. However, the statistical signature of such changes when affecting a large number of gene families can be robustly detected. Our analysis indeed revealed an excess of 3083 gene families with a positive association between GFS and MLSP, a figure well above chance expectations. Moreover, 170 gene families were found to be robustly associated with MLSP after correcting for multiple testing and this result persist once we have account for the shared variance given by the phylogenetic relationships between mammalian species, supporting the view that this variation may be under selection.

Though the genetic contribution to longevity per individual gene may not be large, the combined effect of gene groups have a larger effect. Remarkably, in the present study we found an overrepresentation gene variants previously associated to longevity in humans among the gene families showing GFS variations in line with MLSP. Which denote the importance of certain gene groups that, potentially increasing diversity through different mechanisms can influence longevity. This result strongly indicates a functional role for gene family variations during the evolution of longevity in mammals.

Numerous gene mutations and experimental manipulations have been shown to influence or extend lifespan in a variety of model organisms ranging from yeast to mammals (Johnson et al 2002, Suh et al 2008). It has become increasingly apparent that most of those interventions ultimately interface with cellular stress response mechanisms, suggesting that longevity is intimately related to the ability of the organism to effectively cope with both intrinsic and extrinsic stress. Accordingly in a comparison between human and chimpanzee Perry et al. found an increased copy number variation particularly in inflammation response genes (Perry et al 2008) when the predominantly cause of death in humans are atherosclerosis, diabetes, obesity and neurodegenerative diseases compare to that of adult chimps who died mostly by infections (Finch et al 2010).

Our inspection of biological process gene ontology terms revealed a consistent significant enrichment of genes involved in immune and defence response processes within the families where GFS changes occur along with changes in MLSP. This observation suggests that during evolution of longevity in mammals, the required adaptations of defence response mechanisms were, at least in part, brought about by changes in the number of gene copies within selected gene families. Gene products embedded or attached to the plasma membrane, as well as those components of the intermediate filament, suggesting that the intercommunication with the surrounding environment as well as the spatial integration of the cell components as important factors that could affect lifespan.

The fact that the observed enrichment of enlarged families in line with MLSP is significantly pronounced in certain functionally defined sets but not others, further supports an instrumental role for variations in GFS during evolution of mammalian longevity. Our results support the notion that lifespan-associated GFS variations represent an evolutionary response to the specific functional demands of longer lifespan in mammals.

Conclusion

Using a comparative approach including 28 sequenced mammalian species and over 11000 gene families to look for parallel changes in GFS and maximum lifespan, we

demonstrated that these changes are not secondary to known correlates of lifespan (degree of encephalization and neocortex to brain ratio) and are not completely explained by phylogenetic effects. The functional enrichment analysis, suggest that during evolution of longer lifespan in mammals, the underlying adaptations of immune and defence response mechanisms were, at least in part, brought about by changes in the number of gene copies within selected gene families. Most remarkably these gene families are overrepresented in gene variants previously associated to longevity in humans. We propose that the underlying genetic adaptations for a longer lifespan were in part brought about by changes in the number of gene copies within selected gene families.

Acknowledgments

This study was supported by a PhD CONACYT scholarship for JMS and ACM, a Royal Society Dorothy Hodgkin Research Fellowship (DH071902), Royal Society research grant (RG0870644) and a Royal Society research grant for fellows (RG080272) to AUO and a University of Lincoln PP grant to HG.

Author contributions

All authors contributed to the conception and design of the study. JMS and ACM carried out analyses presented. JMS, AOU and HG wrote the manuscript with contributions from all authors.

Tables

Table 1. Maximum lifespan (MLSP), encephalization index (Ei) and neocortex to brain ratio (Nr) for the 28 mammalian species analysed.

Species name	Common name	MLSP	Ei	Nr
Ailuropoda melanoleuca	Giant Panda	36.8	-2.01376	1.81
Callithrix jacchus	Marmoset	16.5	-1.62664	1.52
Canis familiaris	Dog	24	-1.69931	0.63
Cavia porcellus	Guinea Pig	12	-2.94818	0.51
Echinops telfairi	Tenrec	19	-3.27431	0.1
Erinaceus europaeus	Hedgehog	11.7	-2.86287	0.21
Gorilla gorilla	Gorilla	55.4	-1.41552	2.65
Homo sapiens	Human	122.5	0.151656	4.1
Loxodonta africana	Elephant	65	-1.08197	1.72
Macaca mulatta	Macaque	40	-1.19216	2.6
Macropus eugenii	Wallaby	15.1	-2.20734	0.61
Microcebus murinus	Mouse Lemur	18.2	-1.9849	0.79
Mus musculus	Mouse	4	-2.83246	0.32
Mustela putorius furo	European polecat	11.1	-2.54831	0.87
Ornithorhynchus anatinus	Platypus	22.6	-2.21869	0.92
Ovis aries	Sheep	22.8	-1.96105	1.16
Pan troglodytes	Chimpanzee	59.4	-0.94796	3.22
Papio anubis	Olive baboon	37.5	-1.17774	2.76
Pongo abelii	Orangutan	59	-0.89249	2.6
Procavia capensis	Hyrax	14.8	-2.25494	0.78
Pteropus vampyrus	Megabat	20.9	-2.20381	0.68
Rattus norvegicus	Rat	5	-2.86154	0.52
Sarcophilus harrisii	Tasmanian devil	13	-2.79237	0.33
Sorex araneus	Shrew	3.2	-2.83174	0.16
Sus scrofa	Pig	27	-2.46825	1.04
Tarsius syrichta	Tarsier	16	-1.79503	1.09
Tursiops truncatus	Dolphin	51.6	-0.32137	3.78
Vicugna pacos	Alpaca	25.8	-1.68822	1.28

Figures

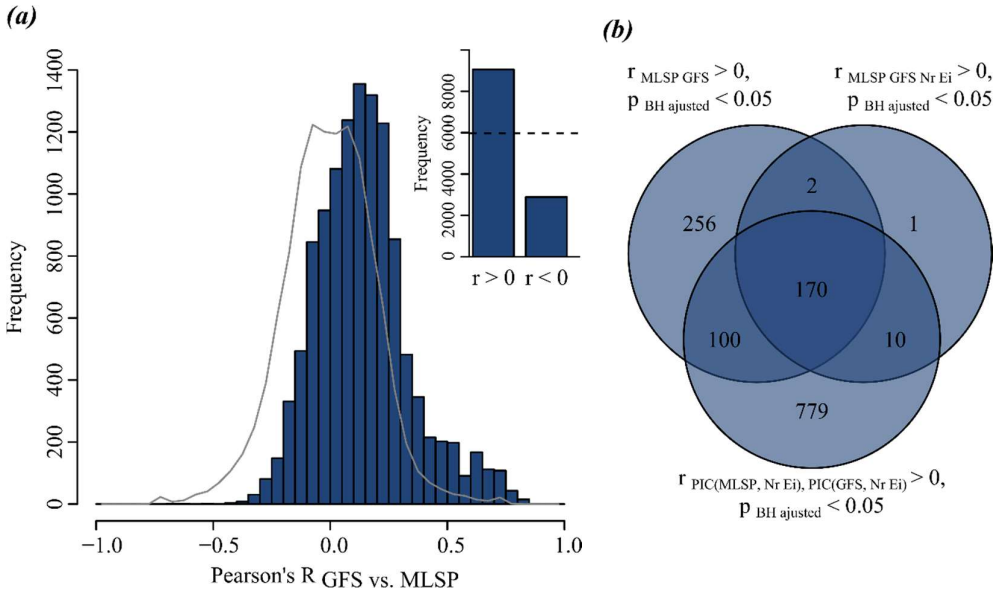


Figure 1. Enrichment of gene family size variations in line with increased lifespan in mammals. (a) Histogram showing the distribution of correlation coefficients for GFS and MLSP in 11943 gene families encompassing 28 mammalian genomes. Expected distribution derived from 10000 Monte Carlo simulations is represented by the solid grey line. Inset: Barplot showing the number of positive and negative correlations relative to the expected frequency (dashed line), a strong bias in the number of positive correlations was found (X^2 = 3184.45, p ≈ 0). (b) Venn Diagram showing the gene families displaying a significant association between GFS and MLSP, obtain with either a simple Pearson correlation, using partial correlations in order to control for Nr and Ei, or through PIC analysis.

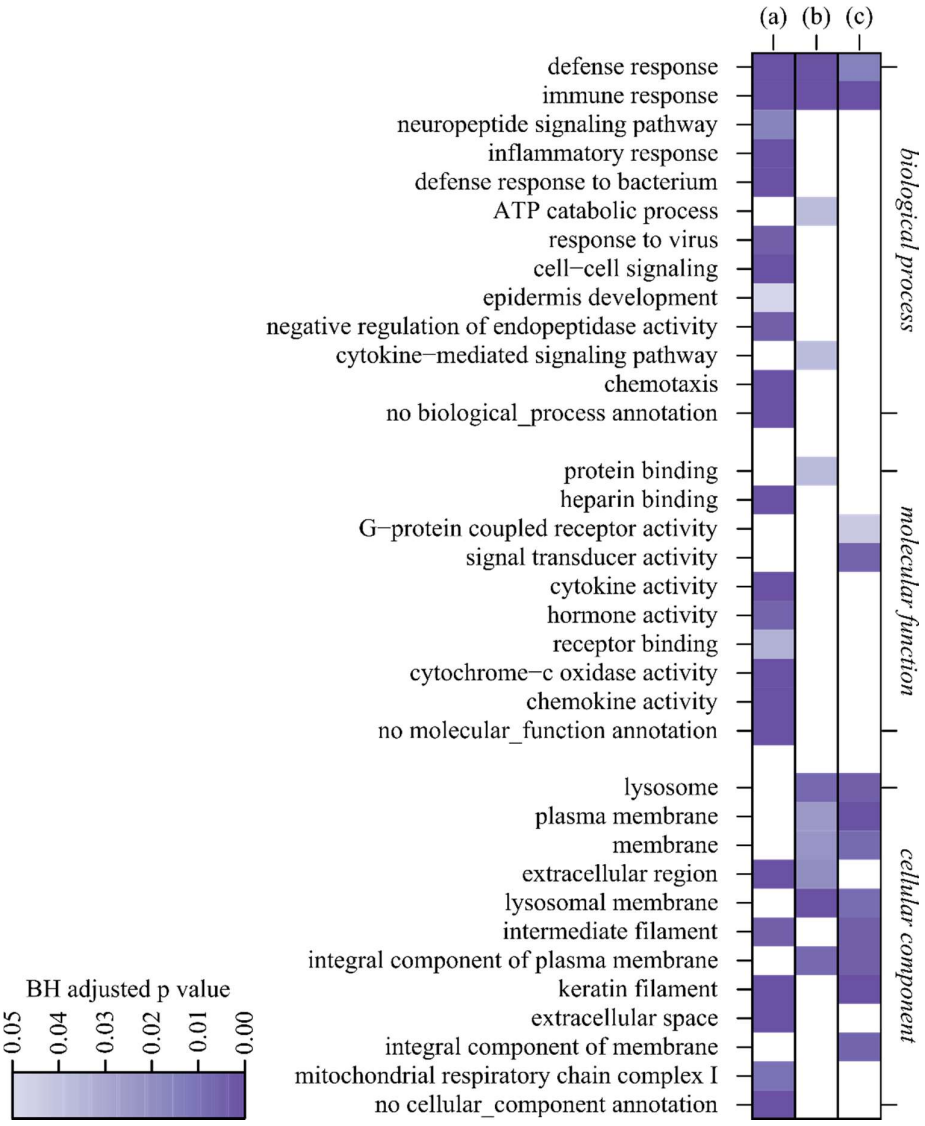


Figure 2. Gene Ontology enrichment analysis of families with GFS variations in line with MLSP. We performed a gene ontology (GO) term enrichment analysis among the families with a significant association between GFS and MLSP, the heatmap shows the Benjamini Hochberg adjusted p value per GO term, as the intensity of the colour increases the adjusted p value is smaller for, white denote GO terms not significantly enriched. Included in the heatmap are only GO terms significantly enriched in (a) the gene families most significantly associated with species maximum lifespan ($r_{MLSP\ GFS} > 0$, $p_{BH\ adjusted} < 0.05$, $n = 528$) (b) the gene families whose GFS variations display the most significant association with MLSP after accounting for the shared variance with Ei and Nr ($r_{MLSP\ GFS\ Nr\ Ei} > 0$, $p_{BH\ adjusted} < 0.05$, $n = 183$) or (c) gene families displaying a significant association between GFS and MLSP once we account for the confounding phenotypes and the phylogenetic relatedness ($r_{PIC\ (GFS\ \sim\ Nr,\ Ei)\ PIC\ (MLSP,\ Nr\ Ei)} > 0$, $p_{BH\ adjusted} < 0.05$, $n = 1059$).

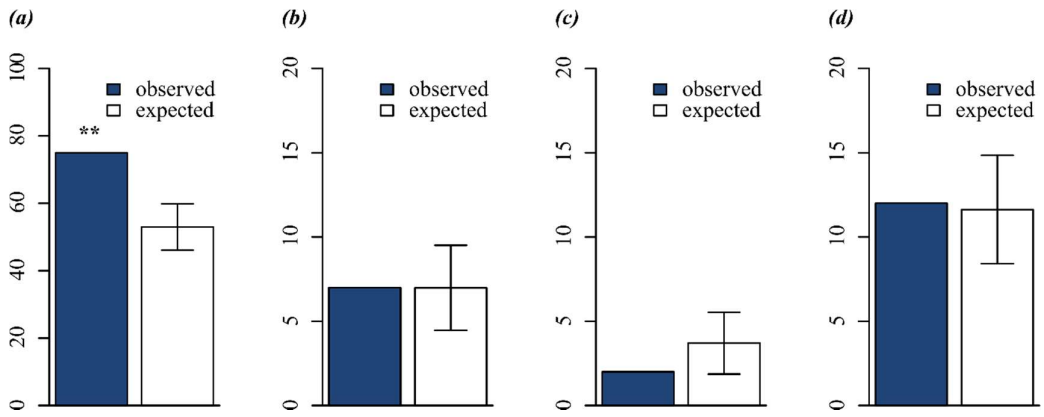


Figure 3. Enrichment analysis of gene compilations regulating, influencing or associated to longevity. We performed an enrichment analysis of genes related to longevity among the families with a significant association between GFS and MLSP. Barplots show the observed and expected number of gene families annotated to ***(a)*** LongevityMap, which is a curated collection of gene variants that had been found associated to longevity in human populations, ***(b)*** GenDR, a set of genes related to dietary restriction known to delay degeneration by ageing and extending lifespan in multiple organisms, ***(c)*** DeathBase, a compilation of genes involved in apoptosis and ***(d)*** REPAIRtoire a set of genes involved in DNA repair. Error bars denote the standard error of the mean from 1000 random samples drawn from the whole population. ** denote p value < 0.001.

General discussion

In this thesis I explored functional associations between components from the genome, transcriptome and to a minor degree the proteome and the phenotypic complexity of the nervous system. I have started by asking whether transcriptional networks (and their underlying regulatory architectures) are a variable or a constant feature of normal development in the brain. The widespread regulatory changes we observe, not only reveals an additional level of complexity in the developmental dynamics but also illustrates the flexibility with which biological components rearrange into different genetic networks to achieve specific functions at particular times. While differential coexpression analysis has been widely used to characterize disease conditions, there has been fewer attempts to characterize changes in the coexpression networks during normal development. Here we provide evidence that such rearrangements occur. We next sought to determine the extent of involvement of immune system-related signalling and regulatory components in the normal development of the nervous system. Our results reveal a potential recruitment of these genes by the nervous system, to achieve neural specific functions. We next, took advantage of the occurrence of coexpression network rearrangements at different developmental stages to examine possible changes in the functional role or the proliferative regulator *yap1* in the context of prenatal brain development to elucidate its involvement in other non-proliferative functions.

Finally, at whole genome level we showed that changes in particular gene families associated to lifespan have occurred preferentially in immune and defence response-related gene families, illustrating the particular importance of the immune system on the evolutionary changes that allowed some mammalian species to increase their lifespan.

Evidence of intense regulatory reorganization occurring in this transition from prenatal to postnatal brain development

Several studies of the human brain transcriptome have shed light into how the different brain structures, hemispheres, ages, gender and even cell types differ in their specific

gene expression patterns (Darmanis et al 2015, Kang et al 2011, Oldham et al 2008, Weickert et al 2009). During the development of the nervous system we need to coordinate the regulation and expression of thousands of genes to ensure the functional integrity of this complex system. Given that most of the genes are at some time and/or place expressed in the human brain (Johnson et al 2009, Kang et al 2011, Stead et al 2006), we expect that a large proportion of the variance in the transcriptome profiles would reflect the particular developmental trajectories of specific tissues.

In the analysis presented in Chapter 1, we used gene expression data for a diversity of cortical areas through a series of developmental stages and found that most of the variation in the transcriptome can be explained by the developmental age and particularly by the division of prenatal and postnatal stages. Even when we incorporated ontogenetically diverse structures such as cerebellum, amygdala and hippocampus (but through a reduced number of developmental ages), the main source of variance in the transcriptome was markedly associated to the transition from prenatal to postnatal development. Although changes in the expression of a particular gene can have a profound phenotypic effect, most of the cellular, physiological and developmental functions are the result of gene groups interacting and cooperating instead of individual genes acting in isolation (Hartwell et al 1999). Clusters of genes involved in the same pathway, biological process, and/or even under common regulators tend to display correlated expression patterns reflecting their functional associations (Eisen et al 1998, Homouz & Kudlicki 2013, Marco et al 2009).

Being the development of the nervous system an extremely dynamic, elaborated and complex process, the precise temporal and special coordination of the gene expression is crucial. In the study described in the first chapter we asked whether transcriptional changes during brain development are just limited to the expression levels of certain genes or gene groups, or alternatively whether these changes are the result of a wider regulatory reorganization affecting the way in which genes are coordinated with each other. In other words, we tested whether the architecture of the underlying regulatory network is a constant or variable feature of the developmental programme. It is important to mention that either model is compatible with our finding of a sharp divide in expression patterns between specific for prenatal and postnatal samples. In the case of the existence of a more dynamic network, this creates a larger number of

possibilities given a limited number of genes, where individual genes have the potential to participate in numerous seemingly unrelated functions depending on which genes they associate with.

Changes in the coexpression of certain gene groups have been associated to pathological states even in the absence of significant changes of gene expression (Choi et al 2005, Miller et al 2008). These changes in gene coexpression effectively reveal events of deregulation and dysfunction affecting specific certain biological processes (i.e. energy metabolism in cancer) or the (pathological) coordination of other genes to promote certain functions (i.e. proliferative functions in cancer) (Choi et al 2005). Changes in coexpression have been described in the context of age-related changes, and as such, mainly linked to aspects of functional decline (Southworth et al 2009).

However changes in coexpression throughout normal development have so far never been explored. In our study, we found a massive rearrangement in the architecture of the global coexpression network in the transition from prenatal to postnatal development. First we demonstrated that even when we divided all expression data into subsets according to brain structures and developmental stage, the resulting gene coexpression networks are more similar between different structures of the same developmental period than they are to themselves before and after birth. Having observed that the coexpression networks group in two defined clusters before and after birth, we used a differential coexpression analysis approach to directly compare regulatory architectures at prenatal and postnatal development (Tesson et al 2010). This method allows unsupervised clustering of differential coexpression networks and is sensible to both changes in gene coexpression within a module between conditions and to changes in module to module coexpression between conditions. Our results revealed a widespread regulatory rearrangement of the entire transcriptome in the perinatal boundary and that this rearrangement is itself organized into discreet reorganization modules.

Moreover the resulting differentially coexpressed gene clusters are functionally coherent, that is, they are integrated by genes preferentially associated to particular biological processes, revealing a reorganization in which each of the clusters consist of genes involved in specific and almost non-overlapping sets of functions.

The fact that this massive rearrangement precisely coincide with the transition from prenatal to postnatal development, may reflect the specific functional demands associated to this abrupt change in environmental conditions. Apart from the obvious alterations that occur at birth, including going from intra to extra uterine conditions, changes in the availability and type of nutrients, levels of oxygen and the exposition to a different array of stressors; the brain goes through an extremely active period during which most neuronal connections are formed as the newborn refine motor and sensory skills, learns and remember from its surrounding, develops social interactions and learns a language. This major transition demands a whole new set of cellular functions required to adapt the developmental programme to the new environmental conditions.

Coordinated pattern of expression of the immune system-related genes in the context of the developing and adult human brain.

In recent years, several signalling and regulatory component of the immune system have been described to play key roles in a variety of neural-specific functions, including critical roles in the developing neural system such as promoting survival of neural precursors, dendritic and axonal growth in addition to the participation in synaptic remodelling and more elaborated cognitive functions such as learning and memory (Carriba et al 2015, Galenkamp et al 2015, Gavalda et al 2009, Gutierrez & Davies 2011, Gutierrez et al 2005, Gutierrez et al 2013, Nolan et al 2011, O'Keeffe et al 2008). Genes displaying a highly variable pattern of expression in the brain during childhood have been found enriched in immune system-associated genes along with other genes such as BDNF which has a critical role during brain development (Sterner et al 2012). Recently, based in a comparative approach we have identified gene family size variations in line with encephalization in mammalian species and found among these gene families an overrepresentation of immune system-related genes (Castillo-Morales et al 2014). Taken together these findings suggest a potentially wider involvement of large numbers of immune system-related genes in key aspects of nervous system development and function.

While the involvement of isolated immune-related signalling components in neural functions may reflect their otherwise ubiquitous character, it could alternatively reflect

a much wider genetic network of immune–related molecules acting as an intrinsic component of the neural-specific regulatory machinery that shapes the functional complexity of the nervous system.

In chapter two, we aimed to gain insights into the wider functional organization of immune-related genetic networks in the developing and adult nervous system and to this end we examined the large scale pattern of coexpression of these genes in the developing and early adult brain.

Our results show a highly significant correlated expression and transcriptional clustering among immune-related genes in the developing and adult brain.

It is worth noting that, to avoid any overestimation of the involvement of the IS genes in the nervous system, we remove from all analyses all the genes that were already annotated to both immune and neurological system process, it is worth mentioning that there is a large overlap between this two gene ontology annotations, over a quarter of immune system genes are also annotated to neurological system process, and even taking these genes out we found a highly coordinated activity of the remaining immune system-related genes in the human brain.

We estimated this coordinated involvement of IS genes in the nervous system as the median coexpression level and compared it to the background population of genes, we did not only found that IS genes have a higher coexpression to other IS genes than expected by chance, but we find that IS genes display an ever higher clustering when compared to random networks of the same density and the equal density and degree distribution. These findings reveal a robust functional association among large numbers of IS components in the nervous system.

In addition we found that this coordinated pattern of expression between IS genes is not a generalized feature of these genes in any tissue. To this end we analysed microarray gene expression data from endothelial cells, liver, muscle, kidney and brain and found that among this tissues it is precisely in the brain where the highest level of coordinated expression is observed.

The fact that immune system genes are highly coexpressed in the developing human brain does not necessarily mean that they are involved in neural specific functions. However, because these analyses were done in normal healthy human tissues, it is unlikely that their highly coordinated expression is related to an immune function or response. Instead, these findings suggest that IS-associated genes are directly engaged in normal neural specific functions. This assertion is further supported by the finding that apart from being highly coexpressed with each other, IS –associated genes are also significantly coexpressed with genes specifically annotated to "nervous system processes". Together these results suggest that IS-associated genes are not only highly coordinated in the nervous system but are also heavily engaged in neural specific functions and further support the notion of a coherent recruitment, by the nervous system, of a substantial proportion of immune system-related signalling and regulatory components.

If, as our results suggest, there is a coherent functional recruitment of entire segments of the IS regulatory machinery by the nervous system, we would expect evidence of this clustering in preparations of dissociated neurons in culture. We experimentally tested this assertion using cultures of dissociated sympathetic neurons. We took advantage of the fact that sympathetic neurons are known to respond to the inflammatory cytokine TNA-alpha (Nolan et al 2014). Accordingly, we stimulated cultured sympathetic neurons with TNF-alpha and characterized the transcriptome by microarray profiling. Among the up-regulated (and also down regulated) genes we found a significant overrepresentation of IS components, (significantly larger than expected by chance). This finding is in agreement with the notion of a modular recruitment of immune system regulatory components by the nervous system. As mentioned in chapter two, the fact that we tested a prediction derived from human expression patterns in experimentally tractable cultured neurons derived from the developing rat, further suggests that the coherent recruitment of IS regulatory clusters is conserved between rodents and humans.

Taken together, our findings support the idea of a widespread and modular recruitment of IS regulatory and signalling circuits by the NS developmental programme in both primates and rodents.

Yap1 function indicated by its coexpressed genes in different brain developmental stages

YAP1 has been involved in a diversity of functions, being proliferation one of the best studied. Its functional diversity is reflected by its involvement in numerous signalling pathways, its direct interaction with hundreds of proteins, including, most importantly, a wide range of transcription factors and consequent its involvement in promoting the expression of a large number of target genes. Given that *yap1* has shown the potential to respond and interact with so many other genes, and given that genes working in similar functions, and pathways are likely to share patterns of expression, we used a differential coexpression approach to identify potential *yap1* interactors at different stages of brain development.

We looked for close coexpression associates of *yap1* at two contrasting stages of brain development: prenatal and postnatal development. It is important to stress that prenatal development is characterized by a highly proliferative activity due to the rapid growth of the brain at this stage. Postnatal brain development, on the other hand, is characterized by a substantially reduced proliferative activity (Bergmann et al 2015, Bhardwaj et al 2006, Goritz & Frisen 2012, Jiang & Nardelli 2015). This contrast between prenatal and postnatal brain development offers an opportunity to identify potentially different roles for *yap1* at these two stages state by identifying existing differences in complement of close associates at these two stages.

While our results reveal a proportion of genes that are constantly coexpressed to *yap1* at both prenatal and postnatal stages, these genes are not the strongest correlates of *yap1*. These observation indicates that those genes that are transiently coexpressed at either stage are not the trivial result of noise in the analysis. Indeed, we find that *yap1* is highly coexpressed with different sets of genes before and after birth. The importance of the specific split of close correlates occurring at the perinatal boundary is underscored by the fact that when we compared with random partition of the samples, none of these reveal significant differences in the complement of coexpression associates of *yap1* strongly suggesting that the observed differences only respond to the existing differences between prenatal and postnatal development.

We propose that the observed differences in *yap1* coexpression profiles specifically respond to existing differences in the functional demands of prenatal and postnatal development, is, therefore, most likely a reflection of the different functions adopted by *yap1* at these two stages. Along these lines, we confirmed that coexpression associates of *yap1* at these two developmental periods are enriched in components located in specific and non-overlapping cellular compartments. At prenatal development most *yap1* associates are enriched in components located in nuclear compartments such as chromosomes and the replication fork. By contrast *yap1* coexpressed genes during postnatal development are cytoplasmic or membrane components (i.e. focal adhesion, cell cortex). Second, functional enrichment analysis aligned with these findings showing that, prenatally, *yap1* associates are enriched in proliferative-related functions (i.e. mitosis, cell cycle) whereas postnatal associates of *yap1* suggest the emergence a complete new set of cellular functions with a statistical enrichment of components involved in processes to detect, integrate signals and engage in a range different cellular response (i.e. migration, differentiation, morphogenesis, metabolic and biosynthetic processes). In line with these observations we found an almost completely different set of upstream regulators of coexpressed associates of *yap1* at these two developmental stages.

Our analysis identifies previously known YAP1 protein-protein interactors, significantly overrepresented in the prenatal period and in agreement with the highly proliferative activity of this developmental period (Hauri et al 2013).

The physiological significance of our findings was supported by *yap1* knockdown experiments in spheroid human cells followed by transcriptome profiling, where we found an overrepresentation of up and down regulated genes the *yap1* coexpressed genes we identify in our analyses.

As for the emerging non proliferative function of *yap1* associates during postnatal development, we found compelling evidence of its potential role in tissue tension at this stage, being tissue tension a new function recently described for *yap1* (Porazinski et al 2015).

Along these lines, among the genes coexpressed with *yap1* during postnatal development, we found *arhgap31* another GTPase activating protein, which has been associated to a rare disease distinguished by scalp and limb defects (Isrie et al 2014). Also coexpressed with *yap1* postnatally was *fn1* and *msn*, major components of the extracellular matrix and a linker between cell cortex and plasma membrane, respectively, and known to be downregulated by miR-200c (Howe et al 2011), a microRNA significantly overrepresented among the same group of *yap1* coexpressed genes at postnatal period. Moreover several proteins responsible for mediating cell-cell contacts were also found. Among these *ocln* and *marveld2* mediate bicellular and tricellular contacts respectively (Ikenouchi et al 2005). Also important players in tissue tension such as *cdh20* cadherin and the cadherin associated protein *ctnna1* (Foty & Steinberg 2005) were found among postnatal *yap1* associates. Furthermore members of the integrin gene family including *itga1*, *itgb1* and *itgb5*, along with *erz*, which forms a complex with *msn* and *vim* are key components incorporating environmental cues to generate a cell response, we suggest that these components may be crucial for a non-proliferative role of *yap1*. Interestingly, *erz* and *vim* are among the differentially expressed genes at *yap1* knockdown in human spheroids. Taken together this evidence suggests the involvement of *yap1* in integrating and presumably maintaining tissue tension at postnatal brain development.

GFS variation as part of the underlying genomic changes associated to lifespan.

There have been a number of comparative studies looking for genomic signatures associated to lifespan. Thus for instance, studies focused in proteins under accelerated evolution at particular lineages of long-lived species have found particularly overrepresented components of the proteasome–ubiquitin system and genes in charge of cellular responses to damage such as COL3A1, DDB1, and CAPNS1 (Li & de Magalhaes 2013).

Gene duplications provide the raw material to increase genetic variability. Over evolutionary time, rounds of gene gain and loss in specific gene families have resulted in differences in the number of genes in different gene families across species. Indeed, previous comparative analyses have suggested that differences in gene family size

could be shaping phenotypic differences among species (Fortna et al 2004). Lifespan is also a highly variable phenotypic trait with differences of orders of magnitude across mammalian species.

In chapter 4 we presented a comparative analysis across 28 mammalian species where we demonstrate a significant overrepresentation of GFS variations in line with increased lifespan. This effect is independent of other phenotypic correlates of lifespan such as encephalization and neocortex to brain ratio. A significant bias in the positive association between increased longevity and gene family size was detected in an excess of over 3000 gene families showing a positive correlation relative to random expectations. This bias remained after correcting for the shared variance accounted for by the phylogenetic relatedness between species with at least 170 gene families robustly associated to maximum lifespan.

It is worth mentioning that we used maximum lifespan (MLSP) as a measure of longevity, given that it better reflect the intrinsic longevity potential (de Magalhaes & Costa 2009).

Interestingly, genes with variants previously associated with longevity in human populations were found overrepresented among MLSP-associated gene families. This observed association between lifespan and gene family size and the enrichment of longevity associated genes among these gene families suggest a functional role for gene family variations during the evolution of longevity in mammals.

Among the genes families displaying a strong correlation with maximum lifespan we found an overrepresentation of gene families involved in immune and defence response. Along this lines, we can speculate that long-lived species are expected to demand more efficient mechanisms to cope both intrinsic and extrinsic stressors and pathogens. Our finding of a particular enrichment of gene families involved immune and defence response, supports the notion that during evolution of longevity in mammals, the required adaptations of defence response mechanisms were, at least in part, brought about by changes in the number of gene copies within selected gene families. Additionally, we found an overrepresentation of particular molecular functions and cellular components that further suggest the importance of sensing the

surrounding environment and signal transduction as key mechanisms to in the evolution of longevity in mammals.

Together, the results presented support the idea that lifespan-associated GFS variations represent an evolutionary response to the specific functional demands of longer lifespan in mammals.

General conclusions

Using a range of functional genomics approaches, including differential expression, as well as coexpression, and differential coexpression network approaches we have explored the dynamics of gene regulatory networks at different levels of nervous system development. First, we described a widespread and modular reorganization of the global network of gene regulatory interactions during perinatal human brain development. Next, we examined the modular and coordinated expression of immune system regulatory and signalling components in the developing and adult nervous system, discussing its functional significance. Then, we focused in the transcriptional co-activator *yap1*, where we were able to identify transient functional associations differentially engaged between prenatal and postnatal brain development suggesting the involvement *yap1* in distinct non overlapping functions at these two developmental stages. Finally, in our last study we used a slightly different approach comparing mammalian species to investigate the genomic bases of organismal lifespan, and proposed that underlying genetic adaptations for a longer lifespan were in part brought about by changes in the number of gene copies of specific gene families.

References

Alarcon C, Zaromytidou AI, Xi Q, Gao S, Yu J, et al. 2009. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell* 139: 757-69

Allen JS, Bruss J, Damasio H. 2005. The aging brain: the cognitive reserve hypothesis and hominid evolution. *American journal of human biology : the official journal of the Human Biology Council* 17: 673-89

Allocco DJ, Kohane IS, Butte AJ. 2004. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics* 5: 18

Amar D, Safer H, Shamir R. 2013. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS computational biology* 9: e1002955

Amieva MR, Furthmayr H. 1995. Subcellular localization of moesin in dynamic filopodia, retraction fibers, and other structures involved in substrate exploration, attachment, and cell-cell contacts. *Experimental cell research* 219: 180-96

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* 11: R106

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-9

Atladottir HO, Thorsen P, Ostergaard L, Schendel DE, Lemcke S, et al. 2010. Maternal infection requiring hospitalization during pregnancy and autism spectrum disorders. *Journal of autism and developmental disorders* 40: 1423-30

Attisano L, Wrana JL. 2013. Signal integration in TGF-beta, WNT, and Hippo pathways. *F1000prime reports* 5: 17

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* 5: 101-13

Barrickman NL, Bastian ML, Isler K, van Schaik CP. 2008. Life history costs and benefits of encephalization: a comparative test using data from long-term studies of primates in the wild. *Journal of human evolution* 54: 568-90

Beatty J, Laughlin RE. 2006. Genomic regulation of natural variation in cortical and noncortical brain volume. *BMC neuroscience* 7: 16

Berchtold NC, Cribbs DH, Coleman PD, Rogers J, Head E, et al. 2008. Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proceedings of the National Academy of Sciences of the United States of America* 105: 15605-10

Bergmann O, Spalding KL, Frisen J. 2015. Adult Neurogenesis in Humans. *Cold Spring Harbor perspectives in biology* 7: a018994

Bhardwaj RD, Curtis MA, Spalding KL, Buchholz BA, Fink D, et al. 2006. Neocortical neurogenesis in humans is restricted to development. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12564-8

Boksa P. 2010. Effects of prenatal infection on brain development and behavior: a review of findings from animal models. *Brain, behavior, and immunity* 24: 881-97

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-20

Brashear A, Ozelius LJ, Sweadner KJ. 2014. ATP1A3 mutations: what is the phenotype? *Neurology* 82: 468-9

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478: 343-8

Breitling R, Armengaud P, Amtmann A, Herzyk P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters* 573: 83-92

Budovsky A, Craig T, Wang J, Tacutu R, Csordas A, et al. 2013. LongevityMap: a database of human genetic variants associated with longevity. *Trends in genetics : TIG* 29: 559-60

Bush EC, Allman JM. 2004. The scaling of frontal cortex in primates and carnivores. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3962-6

Busk M, Pytela R, Sheppard D. 1992. Characterization of the Integrin-Alpha-V-Beta-6 as a Fibronectin-Binding Protein. *Journal of Biological Chemistry* 267: 5790-96

Camargo FD, Gokhale S, Johnnidis JB, Fu D, Bell GW, et al. 2007. YAP1 increases organ size and expands undifferentiated progenitor cells. *Current biology : CB* 17: 2054-60

Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. 2006. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC genomics* 7: 40

Carriba P, Jimenez S, Navarro V, Moreno-Gonzalez I, Barneda-Zahonero B, et al. 2015. Amyloid-beta reduces the expression of neuronal FAIM-L, thereby shifting the inflammatory response mediated by TNFalpha from neuronal protection to death. *Cell death & disease* 6: e1639

Castillo-Morales A, Monzon-Sandoval J, Urrutia AO, Gutierrez H. 2014. Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions. *Proceedings. Biological sciences / The Royal Society* 281: 20132428

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, et al. 2009. Conservation of core gene expression in vertebrate tissues. *Journal of biology* 8: 33

Chen C, Hyun TK, Han X, Feng Z, Li Y, et al. 2014. Coexpression within Integrated Mitochondrial Pathways Reveals Different Networks in Normal and Chemically Treated Transcriptomes. *International journal of genomics* 2014: 452891

Ching T, Huang S, Garmire LX. 2014. Power analysis and sample size estimation for RNA-Seq differential expression. *Rna* 20: 1684-96

Choi JK, Yu U, Yoo OJ, Kim S. 2005. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* 21: 4348-55

Chu JH, Lazarus R, Carey VJ, Raby BA. 2011. Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. *BMC systems biology* 5: 89

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, et al. 2015. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America* 112: 7285-90

de la Fuente A. 2010. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in genetics : TIG* 26: 326-33

de la Fuente A, Bing N, Hoeschele I, Mendes P. 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20: 3565-74

de Magalhaes JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *Journal of evolutionary biology* 22: 1770-4

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. 2006. The evolution of mammalian gene families. *PloS one* 1: e85

Diboun I, Wernisch L, Orengo CA, Koltzenburg M. 2006. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC genomics* 7: 252

Diez J, Walter D, Munoz-Pinedo C, Gabaldon T. 2010. DeathBase: a database on structure, evolution and function of proteins involved in apoptosis and other forms of cell death. *Cell death and differentiation* 17: 735-6

Dobbing J, Sands J. 1973. Quantitative growth and development of human brain. *Archives of disease in childhood* 48: 757-67

Dowell RD. 2011. The similarity of gene expression between human and mouse tissues. *Genome biology* 12: 101

Ebinger P. 1974. A cytoarchitectonic volumetric comparison of brains in wild and domestic sheep. *Zeitschrift fur Anatomie und Entwicklungsgeschichte* 144: 267-302

Ehmer U, Zmoos AF, Auerbach RK, Vaka D, Butte AJ, et al. 2014. Organ size control is dominant over Rb family inactivation to restrict proliferation in vivo. *Cell reports* 8: 371-81

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95: 14863-8

Fan R, Kim NG, Gumbiner BM. 2013. Regulation of Hippo pathway by mitogenic growth factors via phosphoinositide 3-kinase and phosphoinositide-dependent kinase-1. *Proceedings of the National Academy of Sciences of the United States of America* 110: 2569-74

Felsenstein J. 1985. Phylogenies and the Comparative Method. *The American Naturalist* 125: 1-15

Finch CE, Morgan TE, Longo VD, de Magalhaes JP. 2010. Cell resilience in species life spans: a link to inflammation? *Aging cell* 9: 519-26

Flicek P, Amode MR, Barrell D, Beal K, Billis K, et al. 2014. Ensembl 2014. *Nucleic acids research* 42: D749-55

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, et al. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS biology* 2: E207

Foty RA, Pfleger CM, Forgacs G, Steinberg MS. 1996. Surface tensions of embryonic tissues predict their mutual envelopment behavior. *Development* 122: 1611-20

Foty RA, Steinberg MS. 2005. The differential adhesion hypothesis: a direct evaluation. *Developmental biology* 278: 255-63

Freckleton RP, Harvey PH, Pagel M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat* 160: 712-26

Gaiteri C, Ding Y, French B, Tseng GC, Sibille E. 2014. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, brain, and behavior* 13: 13-24

Gaiteri C, Sibille E. 2011. Differentially expressed genes in major depression reside on the periphery of resilient gene coexpression networks. *Frontiers in neuroscience* 5: 95

Galenkamp KM, Carriba P, Urresti J, Planells-Ferrer L, Coccia E, et al. 2015. TNFalpha sensitizes neuroblastoma cells to FasL-, cisplatin- and etoposide-induced cell death by NF-kappaB-mediated expression of Fas. *Molecular cancer* 14: 62

Garay PA, Hsiao EY, Patterson PH, McAllister AK. 2013. Maternal immune activation causes age- and region-specific changes in brain cytokines in offspring throughout development. *Brain, behavior, and immunity* 31: 54-68

Garbett KA, Hsiao EY, Kalman S, Patterson PH, Mirnics K. 2012. Effects of maternal immune activation on gene expression patterns in the fetal brain. *Translational psychiatry* 2: e98

Gargalovic PS, Imura M, Zhang B, Gharavi NM, Clark MJ, et al. 2006. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proceedings of the National Academy of Sciences of the United States of America* 103: 12741-6

Gary R, Bretscher A. 1993. Heterotypic and homotypic associations between ezrin and moesin, two putative membrane-cytoskeletal linking proteins. *Proceedings of the National Academy of Sciences of the United States of America* 90: 10846-50

Gavalda N, Gutierrez H, Davies AM. 2009. Developmental regulation of sensory neurite growth by the tumor necrosis factor superfamily member LIGHT. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29: 1599-607

Georg Baron HS, Heiko D. Frahm 1996. *Comparative Neurobiology in Chiroptera: Macromorphology, Brain Structures, Tables, and Atlases.*

Gillis J, Pavlidis P. 2009. A methodology for the analysis of differential coexpression across the human lifespan. *BMC bioinformatics* 10: 306

Glebova NO, Ginty DD. 2005. Growth and survival signals controlling sympathetic nervous system development. *Annual review of neuroscience* 28: 191-222

Gonzalez-Lagos C, Sol D, Reader SM. 2010. Large-brained mammals live longer. *Journal of evolutionary biology* 23: 1064-74

Goritz C, Frisen J. 2012. Neural stem cells and neurogenesis in the adult. *Cell stem cell* 10: 657-9

Guerout N, Li X, Barnabe-Heider F. 2014. Cell fate control in the developing central nervous system. *Experimental cell research* 321: 77-83

Gutierrez H, Castillo A, Monzon J, Urrutia AO. 2011. Protein amino acid composition: a genomic signature of encephalization in mammals. *PloS one* 6: e27261

Gutierrez H, Davies AM. 2011. Regulation of neural process growth, elaboration and structural plasticity by NF-kappaB. *Trends in neurosciences* 34: 316-25

Gutierrez H, Hale VA, Dolcet X, Davies A. 2005. NF-kappaB signalling regulates the growth of neural processes in the developing PNS and CNS. *Development* 132: 1713-26

Gutierrez H, Kisiswa L, O'Keeffe GW, Smithen MJ, Wyatt S, Davies AM. 2013. Regulation of neurite growth by tumour necrosis superfamily member RANKL. *Open biology* 3: 120150

Hahn MW, Demuth JP, Han SG. 2007a. Accelerated rate of gene gain and loss in primates. *Genetics* 177: 1941-9

Hahn MW, Han MV, Han SG. 2007b. Gene family evolution across 12 Drosophila genomes. *PLoS genetics* 3: e197

Hakeem AY, Hof PR, Sherwood CC, Switzer RC, 3rd, Rasmussen LE, Allman JM. 2005. Brain of the African elephant (Loxodonta africana): neuroanatomy from magnetic resonance images. *The anatomical record. Part A, Discoveries in molecular, cellular, and evolutionary biology* 287: 1117-27

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. *Genome research* 19: 859-67

Harper JM, Salmon AB, Leiser SF, Galecki AT, Miller RA. 2007. Skin-derived fibroblasts from long-lived species are resistant to some, but not all, lethal stresses and to the mitochondrial inhibitor rotenone. *Aging cell* 6: 1-13

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* 32: D258-61

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature* 402: C47-52

Hauri S, Wepf A, van Drogen A, Varjosalo M, Tapon N, et al. 2013. Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1. *Molecular systems biology* 9: 713

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, et al. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489: 391-9

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC evolutionary biology* 4: 2

Helsmoortel C, Vulto-van Silfhout AT, Coe BP, Vandeweyer G, Rooms L, et al. 2014. A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nature genetics* 46: 380-4

Hergovich A. 2012. Mammalian Hippo signalling: a kinase network regulated by protein-protein interactions. *Biochemical Society transactions* 40: 124-8
Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, et al. 2007. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome biology* 8: R76
Heyer LJ, Kruglyak S, Yooseph S. 1999. Exploring expression data: identification and analysis of coexpressed genes. *Genome research* 9: 1106-15
Hobert O. 2008. Gene regulation by transcription factors and microRNAs. *Science* 319: 1785-6
Holland PW, Garcia-Fernandez J. 1996. Hox genes and chordate evolution. *Developmental biology* 173: 382-95
Homouz D, Kudlicki AS. 2013. The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. *PloS one* 8: e54699
Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. 2006. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 22: 2825-7
Howe EN, Cochrane DR, Richer JK. 2011. Targets of miR-200c mediate suppression of cell motility and anoikis resistance. *Breast cancer research : BCR* 13: R45
http://geneontology.org/. Gene Ontology.
http://www.brainspan.org/. Brainspan database.
http://www.ensembl.org/biomart/. Ensembl Biomart.
http://www.ensembl.org/index.html. Ensembl database.
http://www.ncbi.nlm.nih.gov/geo/. Gene Expression Omnibus.
Hughes AL, Friedman R. 2004. Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. *Journal of molecular evolution* 59: 827-33
Huttenlocher PR, Dabholkar AS. 1997. Regional differences in synaptogenesis in human cerebral cortex. *The Journal of comparative neurology* 387: 167-78
Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S. 2012. Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics* 28: 1592-7
Iannitti T, Palmieri B. 2011. Inflammation and genetics: an insight in the centenarian model. *Human biology* 83: 531-59
Iglesias-Bexiga M, Castillo F, Cobos ES, Oka T, Sudol M, Luque I. 2015. WW domains of the yes-kinase-associated-protein (YAP) transcriptional regulator behave as independent units with different binding preferences for PPxY motif-containing ligands. *PloS one* 10: e0113828
Ikenouchi J, Furuse M, Furuse K, Sasaki H, Tsukita S, Tsukita S. 2005. Tricellulin constitutes a novel barrier at tricellular contacts of epithelial cells. *The Journal of cell biology* 171: 939-45
Isrie M, Wuyts W, Van Esch H, Devriendt K. 2014. Isolated terminal limb reduction defects: extending the clinical spectrum of Adams-Oliver syndrome and ARHGAP31 mutations. *American journal of medical genetics. Part A* 164A: 1576-9
Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411: 41-2
Jiang X, Nardelli J. 2015. Cellular and molecular introduction to brain development. *Neurobiology of disease*
Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, et al. 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 62: 494-509
Johnson TE, Henderson S, Murakami S, de Castro E, de Castro SH, et al. 2002. Longevity genes in the nematode Caenorhabditis elegans also mediate increased resistance to stress and prevent disease. *Journal of inherited metabolic disease* 25: 197-206

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* 478: 483-9

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, et al. 2015. Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. *Science* 348: 1139-43

Kapoor A, Yao W, Ying H, Hua S, Liewen A, et al. 2014. Yap1 activation enables bypass of oncogenic Kras addiction in pancreatic cancer. *Cell* 158: 185-97

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14: R36

Kim SE, Mori R, Komatsu T, Chiba T, Hayashi H, et al. 2015. Upregulation of cytochrome c oxidase subunit 6b1 (Cox6b1) and formation of mitochondrial supercomplexes: implication of Cox6b1 in the effect of calorie restriction. *Age* 37: 9787

Kirkwood TL, Kapahi P, Shanley DP. 2000. Evolution, stress, and longevity. *Journal of anatomy* 197 Pt 4: 587-90

Kisiswa L, Osorio C, Erice C, Vizard T, Wyatt S, Davies AM. 2013. TNFalpha reverse signaling promotes sympathetic axon growth and target innervation. *Nature neuroscience* 16: 865-73

Komuro A, Nagai M, Navin NE, Sudol M. 2003. WW domain-containing protein YAP associates with ErbB-4 and acts as a co-transcriptional activator for the carboxyl-terminal fragment of ErbB-4 that translocates to the nucleus. *The Journal of biological chemistry* 278: 33334-41

Kostka D, Spang R. 2004. Finding disease specific alterations in the co-expression of genes. *Bioinformatics* 20 Suppl 1: i194-9

Kourtis N, Tavernarakis N. 2011. Cellular stress response pathways and ageing: intricate molecular relationships. *The EMBO journal* 30: 2520-31

Kruska D, Rohrs M. 1974. Comparative--quantitative investigations on brains of feral pigs from the Galapagos Islands and of European domestic pigs. *Zeitschrift fur Anatomie und Entwicklungsgeschichte* 144: 61-73

Kruska DCT. 2014. Comparative quantitative investigations on brains of wild cavies (Cavia aperea) and guinea pigs (Cavia aperea f. porcellus). A contribution to size changes of CNS structures due to domestication. *Mamm Biol* 79: 230-39

Kumar S, Hedges SB. 2011. TimeTree2: species divergence times on the iPhone. *Bioinformatics* 27: 2023-4

Lambert N, Lambot MA, Bilheu A, Albert V, Englert Y, et al. 2011. Genes expressed in specific areas of the human fetal cerebral cortex display distinct patterns of evolution. *PloS one* 6: e17753

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 9: 559

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. 2013. Software for computing and annotating genomic ranges. *PLoS computational biology* 9: e1003118

Lecuit T, Lenne PF. 2007. Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis. *Nature reviews. Molecular cell biology* 8: 633-44

LeDoux MS, Xu L, Xiao J, Ferrell B, Menkes DL, Homayouni R. 2006. Murine central and peripheral nervous system transcriptomes: comparative gene expression. *Brain research* 1107: 24-41

Levy D, Adamovich Y, Reuven N, Shaul Y. 2007. The Yes-associated protein 1 stabilizes p73 by preventing Itch-mediated ubiquitination of p73. *Cell death and differentiation* 14: 743-51

Li Y, de Magalhaes JP. 2013. Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age* 35: 301-14

Li Y, Hibbs MA, Gard AL, Shylo NA, Yun K. 2012. Genome-wide analysis of N1ICD/RBPJ targets in vivo reveals direct transcriptional regulation of Wnt, SHH, and hippo pathway effectors by Notch1. *Stem cells* 30: 741-52

Lian I, Kim J, Okazawa H, Zhao J, Zhao B, et al. 2010. The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Genes & development* 24: 1106-18

Liang Y-H, Cai B, Chen F, Wang G, Wang M, et al. 2014. Construction and validation of a gene co-expression network in grapevine (Vitis vinifera. L.). *Horticulture Research* 1: 14040

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, et al. 2013. Global epigenomic reconfiguration during mammalian brain development. *Science* 341: 1237905

Liu CT, Yuan S, Li KC. 2009. Patterns of co-expression for protein complexes by size in Saccharomyces cerevisiae. *Nucleic acids research* 37: 526-32

Low BC, Pan CQ, Shivashankar GV, Bershadsky A, Sudol M, Sheetz M. 2014. YAP/TAZ as mechanosensors and mechanotransducers in regulating organ size and tumor growth. *FEBS letters* 588: 2663-70

Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. 2010. DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one* 5: e9803

Maier T, Guell M, Serrano L. 2009. Correlation of mRNA and protein in complex biological samples. *FEBS letters* 583: 3966-73

Marco A, Konikoff C, Karr TL, Kumar S. 2009. Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics* 25: 2473-7

Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, et al. 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* 348: 660-5

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338: 1593-9

Milanowska K, Krwawicz J, Papaj G, Kosinski J, Poleszak K, et al. 2011. REPAIRtoire--a database of DNA repair pathways. *Nucleic acids research* 39: D788-92

Miller JA, Oldham MC, Geschwind DH. 2008. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 28: 1410-20

Mitchell RH, Goldstein BI. 2014. Inflammation in children and adolescents with neuropsychiatric disorders: a systematic review. *Journal of the American Academy of Child and Adolescent Psychiatry* 53: 274-96

Nikolayeva O, Robinson MD. 2014. edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. *Methods in molecular biology* 1150: 45-79

Ning K, Ng HK, Srihari S, Leong HW, Nesvizhskii AI. 2010. Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. *BMC bioinformatics* 11: 505

Nolan AM, Collins LM, Wyatt SL, Gutierrez H, O'Keeffe GW. 2014. The neurite growth inhibitory effects of soluble TNFalpha on developing sympathetic neurons are dependent on developmental age. *Differentiation; research in biological diversity* 88: 124-30

Nolan AM, Nolan YM, O'Keeffe GW. 2011. IL-1beta inhibits axonal growth of developing sympathetic neurons. *Molecular and cellular neurosciences* 48: 142-50

O'Keeffe GW, Gutierrez H, Pandolfi PP, Riccardi C, Davies AM. 2008. NGF-promoted axon growth and target innervation requires GITRL-GITR signaling. *Nature neuroscience* 11: 135-42

Obayashi T, Kinoshita K. 2011. COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic acids research* 39: D1016-22

Oka T, Mazack V, Sudol M. 2008. Mst2 and Lats kinases regulate apoptotic function of Yes kinase-associated protein (YAP). *The Journal of biological chemistry* 283: 27534-46
Oldham MC, Horvath S, Geschwind DH. 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences of the United States of America* 103: 17973-8
Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, et al. 2008. Functional organization of the transcriptome in human brain. *Nature neuroscience* 11: 1271-82
Orike N, Thrasivoulou C, Cowen T. 2001. Serum-free culture of dissociated, purified adult and aged sympathetic neurons and quantitative assays of growth and survival. *Journal of neuroscience methods* 106: 153-60
Orr BA, Bai H, Odia Y, Jain D, Anders RA, Eberhart CG. 2011. Yes-associated protein 1 is widely expressed in human brain tumors and promotes glioblastoma growth. *Journal of neuropathology and experimental neurology* 70: 568-77
Ousman SS, Kubes P. 2012. Immune surveillance in the central nervous system. *Nature neuroscience* 15: 1096-101
Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401: 877-84
Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20: 289-90
Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, et al. 2013. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155: 1008-21
Perez VI, Buffenstein R, Masamsetti V, Leonard S, Salmon AB, et al. 2009. Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat. *Proceedings of the National Academy of Sciences of the United States of America* 106: 3059-64
Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, et al. 2008. Copy number variation and evolution in humans and chimpanzees. *Genome research* 18: 1698-710
Petrosillo G, De Benedictis V, Ruggiero FM, Paradies G. 2013. Decline in cytochrome c oxidase activity in rat-brain mitochondria with aging. Role of peroxidized cardiolipin and beneficial effect of melatonin. *Journal of bioenergetics and biomembranes* 45: 431-40
Pinna A, Heise S, Flassig RJ, de la Fuente A, Klamt S. 2013. Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation. *BMC systems biology* 7: 73
Pinna A, Soranzo N, de la Fuente A. 2010. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PloS one* 5: e12912
Pirlot P. 1981. A quantitative approach to the marsupial brain in an eco-ethological perspective. *Revue canadienne de biologie / editee par l'Universite de Montreal* 40: 229-50
Pirlot P, Jiao SS. 1985. Quantitative morphology of the panda brain in comparison with the brains of the raccoon and the bear. *Journal fur Hirnforschung* 26: 17-22
Porazinski S, Wang H, Asaoka Y, Behrndt M, Miyamoto T, et al. 2015. YAP is essential for tissue tension to ensure vertebrate 3D body shape. *Nature* 521: 217-21
Rakic P. 2002. Pre- and post-developmental neurogenesis in primates. *Clin Neurosci Res* 2: 29-39
Ransohoff RM, Brown MA. 2012. Innate immunity in the central nervous system. *The Journal of clinical investigation* 122: 1164-71
Reep RL, Finlay BL, Darlington RB. 2007. The limbic system in Mammalian brain evolution. *Brain, behavior and evolution* 70: 57-70

Ricklefs RE. 2010. Life-history connections to rates of aging in terrestrial vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* 107: 10314-9

Ricklefs RE, Cadena CD. 2008. Heritability of longevity in captive populations of nondomesticated mammals and birds. *The journals of gerontology. Series A, Biological sciences and medical sciences* 63: 435-46

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43: e47

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139-40

Rosenbluh J, Nijhawan D, Cox AG, Li X, Neal JT, et al. 2012. beta-Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. *Cell* 151: 1457-73

Rowitch DH, Kriegstein AR. 2010. Developmental genetics of vertebrate glial-cell specification. *Nature* 468: 214-22

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. 2000. Comparative genomics of the eukaryotes. *Science* 287: 2204-15

Saris CG, Horvath S, van Vught PW, van Es MA, Blauw HM, et al. 2009. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC genomics* 10: 405

Schumacher B, Garinis GA, Hoeijmakers JH. 2008. Age to survive: DNA damage and aging. *Trends in genetics : TIG* 24: 77-85

Shen L, Liu G, Zou Y, Zhou Z, Su Z, Gu X. 2015. The evolutionary panorama of organ-specifically expressed or repressed orthologous genes in nine vertebrate species. *PloS one* 10: e0116872

Shortman K, Liu YJ. 2002. Mouse and human dendritic cell subtypes. *Nature reviews. Immunology* 2: 151-61

Smith RP, Lerch-Haner JK, Pardinas JR, Buchser WJ, Bixby JL, Lemmon VP. 2011. Transcriptional profiling of intrinsic PNS factors in the postnatal mouse. *Molecular and cellular neurosciences* 46: 32-44

Sohal RS, Sohal BH, Orr WC. 1995. Mitochondrial superoxide and hydrogen peroxide generation, protein oxidative damage, and longevity in different species of flies. *Free radical biology & medicine* 19: 499-504

Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics* 14: 91

Song L, Langfelder P, Horvath S. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* 13: 328

Sorensen HJ, Mortensen EL, Reinisch JM, Mednick SA. 2009. Association between prenatal exposure to bacterial infection and risk of schizophrenia. *Schizophrenia bulletin* 35: 631-7

Soshnikova N, Dewaele R, Janvier P, Krumlauf R, Duboule D. 2013. Duplications of hox gene clusters and the emergence of vertebrates. *Developmental biology* 378: 194-9

Southworth LK, Owen AB, Kim SK. 2009. Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS genetics* 5: e1000776

Stead JD, Neal C, Meng F, Wang Y, Evans S, et al. 2006. Transcriptional profiling of the developing rat brain reveals that the most dramatic regional differentiation in gene expression occurs postpartum. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 26: 345-53

Stephan H, Frahm H, Baron G. 1981. New and revised data on volumes of brain structures in insectivores and primates. *Folia primatologica; international journal of primatology* 35: 1-29

Sterner KN, Weckle A, Chugani HT, Tarca AL, Sherwood CC, et al. 2012. Dynamic gene expression in the human cerebral cortex distinguishes children from adults. *PloS one* 7: e37714

Strano S, Munarriz E, Rossi M, Castagnoli L, Shaul Y, et al. 2001. Physical interaction with Yes-associated protein enhances p73 transcriptional activity. *The Journal of biological chemistry* 276: 15164-73

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249-55

Sudol M, Shields DC, Farooq A. 2012. Structures of YAP protein domains reveal promising targets for development of new cancer drugs. *Seminars in cell & developmental biology* 23: 827-33

Suh Y, Atzmon G, Cho MO, Hwang D, Liu B, et al. 2008. Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proceedings of the National Academy of Sciences of the United States of America* 105: 3438-42

Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, et al. 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic acids research* 41: D1027-33

Tesson BM, Breitling R, Jansen RC. 2010. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics* 11: 497

Torkamani A, Dean B, Schork NJ, Thomas EA. 2010. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome research* 20: 403-12

Twohig JP, Cuff SM, Yong AA, Wang EC. 2011. The role of tumor necrosis factor receptor superfamily members in mammalian brain development, function and homeostasis. *Reviews in the neurosciences* 22: 509-33

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, et al. 2009. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, cell & environment* 32: 1633-51

Walker RF. 2011. Developmental theory of aging revisited: focus on causal and mechanistic links between development and senescence. *Rejuvenation research* 14: 429-36

Walley AJ, Jacobson P, Falchi M, Bottolo L, Andersson JC, et al. 2012. Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. *International journal of obesity* 36: 137-47

Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research* 41: W77-83

Wang Z, Ye J, Deng Y, Yan Z, Denduluri S, He TC. 2014. Wnt Hippo: A balanced act or dynamic duo? *Genes & Diseases* 1: 127-28

Watts DJ, Strogatz SH. 1998. Collective dynamics of 'small-world' networks. *Nature* 393: 440-2

Weickert CS, Elashoff M, Richards AB, Sinclair D, Bahn S, et al. 2009. Transcriptome analysis of male-female differences in prefrontal cortical development. *Molecular psychiatry* 14: 558-61

Wensink MJ, van Heemst D, Rozing MP, Westendorp RG. 2012. The maintenance gap: a new theoretical perspective on the evolution of aging. *Biogerontology* 13: 197-201

Williams EJ, Bowles DJ. 2004. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome research* 14: 1060-7

Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, et al. 2013. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155: 997-1007

Wolfe CJ, Kohane IS, Butte AJ. 2005. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics* 6: 227

Wuttke D, Connor R, Vora C, Craig T, Li Y, et al. 2012. Dissecting the gene network of dietary restriction to identify evolutionarily conserved pathways and new functional genes. *PLoS genetics* 8: e1002834

Yip AM, Horvath S. 2007. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics* 8: 22

Young JM, Trask BJ. 2002. The sense of smell: genomics of vertebrate odorant receptors. *Human molecular genetics* 11: 1153-60

Yu H, Luscombe NM, Qian J, Gerstein M. 2003. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in genetics : TIG* 19: 422-7

Zaidi SK, Sullivan AJ, Medina R, Ito Y, van Wijnen AJ, et al. 2004. Tyrosine phosphorylation controls Runx2-mediated subnuclear targeting of YAP to repress transcription. *The EMBO journal* 23: 790-9

Zhang B, Kirov S, Snoddy J. 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research* 33: W741-8

Zhang J, Lu K, Xiang Y, Islam M, Kotian S, et al. 2012. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS computational biology* 8: e1002656

Zhang S. 2012. Hierarchical modular structure identification with its applications in gene coexpression networks. *TheScientificWorldJournal* 2012: 523706

Zhao B, Li L, Lu Q, Wang LH, Liu CY, et al. 2011. Angiomotin is a novel Hippo pathway component that inhibits YAP oncoprotein. *Genes & development* 25: 51-63

Zhao B, Li L, Wang L, Wang CY, Yu J, Guan KL. 2012. Cell detachment activates the Hippo pathway via cytoskeleton reorganization to induce anoikis. *Genes & development* 26: 54-68

Zhao B, Wei X, Li W, Udan RS, Yang Q, et al. 2007. Inactivation of YAP oncoprotein by the Hippo pathway is involved in cell contact inhibition and tissue growth control. *Genes & development* 21: 2747-61

Zhao B, Ye X, Yu J, Li L, Li W, et al. 2008. TEAD mediates YAP-dependent gene induction and growth control. *Genes & development* 22: 1962-71

Zhou X, Kao MC, Wong WH. 2002. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America* 99: 12783-8

Appendices

Attached to this thesis there is a published article where I significantly contributed to both experimental design and data analysis.

PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Castillo-Morales A, Monzón-Sandoval J, Urrutia AO, Gutiérrez H. 2014 Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions. *Proc. R. Soc. B* **281**: 20132428. http://dx.doi.org/10.1098/rspb.2013.2428

Received: 17 September 2013
Accepted: 5 November 2013

Subject Areas:
neuroscience, genomics, evolution

Keywords:
encephalization index, brain evolution, gene expression

Authors for correspondence:
Araxi O. Urrutia
e-mail: a.urrutia@bath.ac.uk
Humberto Gutiérrez
e-mail: hgutierrez@lincoln.ac.uk

†These authors contributed equally to this study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2013.2428 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing

Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions

Atahualpa Castillo-Morales[1,†], Jimena Monzón-Sandoval[1,†], Araxi O. Urrutia[1] and Humberto Gutiérrez[2]

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK
[2]School of Life Sciences, University of Lincoln, Lincoln LN6 7TS, UK

Genomic determinants underlying increased encephalization across mammalian lineages are unknown. Whole genome comparisons have revealed large and frequent changes in the size of gene families, and it has been proposed that these variations could play a major role in shaping morphological and physiological differences among species. Using a genome-wide comparative approach, we examined changes in gene family size (GFS) and degree of encephalization in 39 fully sequenced mammalian species and found a significant over-representation of GFS variations in line with increased encephalization in mammals. We found that this relationship is not accounted for by known correlates of brain size such as maximum lifespan or body size and is not explained by phylogenetic relatedness. Genes involved in chemotaxis, immune regulation and cell signalling-related functions are significantly over-represented among those gene families most highly correlated with encephalization. Genes within these families are prominently expressed in the human brain, particularly the cortex, and organized in co-expression modules that display distinct temporal patterns of expression in the developing cortex. Our results suggest that changes in GFS associated with encephalization represent an evolutionary response to the specific functional requirements underlying increased brain size in mammals.

1. Introduction

Mammalian species in general tend to have larger brain to body size ratios compared with other vertebrates and in some primate and cetacean species this relationship is particularly pronounced [1]. Large brains represent an evolutionarily costly adaptation as they are metabolically expensive, demand higher parental investment than in species with smaller brains and impose a substantial delay in reproductive age [1–5]. In spite of the cost and adaptive impact of larger brains, the precise nature of genomic changes accounting for variations in encephalization across mammalian species is at present poorly understood [6,7].

Whole-genome sequencing efforts have made it possible to study not just individual variations in specific sequences, but also large-scale differences in gene complements between species. Although overall gene number has changed little over the past 800 million years of metazoan evolution, comparative genomic studies have found large disparities among organisms in the number of copies of genes involved in a variety of cellular and developmental processes, and analyses of gene family evolution have shown that instances of gene family expansion and contraction are frequent [8–12]. In a recent analysis of *Drosophila* species, for instance, large numbers of gains and losses have been described, with over 40% of all gene families differing in size among the analysed species. Importantly, the fact that, in these species, rapid gene family size (GFS) evolution is accentuated

in some functional categories strongly suggests that changes in gene number within gene families may reflect evolutionary responses to specific adaptive demands [10]. In this regard, gene duplication events specifically linked to distinct aspects of vertebrate evolution have been described. Examples include the expansion, during early evolution of the vertebrate lineage, of HOX and PAX gene families which are widely believed to have played a key part in the evolution of many known vertebrate innovations [13,14].

A major goal in evolutionary neurobiology is to understand the molecular changes underlying the extraordinary expansion in brain size observed in mammalian evolution. Whether changes in the number of copies of genes involved in distinct cellular and developmental functions has contributed to shaping the morphological, physiological and metabolic machinery supporting brain evolution in mammalian lineages is not known.

By conducting a genome-wide analysis of 39 fully sequenced mammalian species, we set out to establish whether changes in GFS can be linked to increased encephalization. Our results reveal a proportion of gene families displaying a positive association between GFS and level of encephalization significantly larger than expected by chance. This bias occurs most prominently in families associated with specific biological functions. By examining expression data in human tissues, we further found that gene families displaying the highest association between encephalization and GFS are also statistically enriched in genes that are prominently expressed in the brain, with maximal expression in the cortex and displaying an expression signature distinctly associated with cortical development.

2. Methods

(a) Gene family annotations

Annotated gene families encompassing 39 fully sequenced mammalian genomes were obtained from Ensembl [15]. In the context of this annotation, a given gene family constitutes a group of related genes that include both paralogues within the same species and orthologues and paralogues from other species. Any given gene can only be assigned to a single gene family. GFS represents the total number of genes per gene family. In order to maximize the number of families covered in this study (more than 10 000), we included all gene families with members present in no less than six of the 39 mammalian species.

(b) Encephalization index

Because larger species have larger brains, it is necessary to estimate brain mass controlling for the allometric effect of body size. We therefore adopted residuals of a log–log least-squares linear regression of brain mass against body mass as this is the most widely accepted index of encephalization (*Ei*; electronic supplementary material, table S1) [16,17]. While direct estimates of the ratio of brain mass to body mass have also been used as an alternative encephalization index [2,18], this measure is known to be poorly related to brain complexity across taxa [16,17]. Accurate estimates of brain residuals based on a sample of 493 mammalian species were kindly provided by Gonzalez-Lagos *et al.* [2].

(c) Correlation coefficients of gene family size and encephalization index

Simple Pearson correlations between *Ei* and GFS as well as multiple regressions (where maximum lifespan (MLSP) was included as covariate, see below) were carried out using R-based statistical functions. Numerical randomizations to determine statistical significance were conducted using specially written R-based scripts.

(d) Gene ontology terms analysis

Gene ontology (GO) annotations were obtained from the Gene Ontology database (www.geneontology.org). In this study, a particular GO term was associated with a family whenever that term was linked to any of its members in any species. Only terms found to be linked with more than 50 families were examined.

For each GO category, the average Pearson correlation coefficient was calculated. Statistical significance and expected average Pearson correlation per GO was measured using at least 10 000 equally sized random samples taken from the whole gene family population to directly determine the corresponding *p*-values. Bonferroni correction was used in all analyses to correct for multiple tests.

Enrichment analysis of GO categories was carried out by counting the number of families assigned to each GO term within the analysed set of gene families. However, any bias in family counts per GO within a set of families could be owing to a bias in the overall density of GO annotation events within that sample. In order to adjust for differences in the density of GO annotations between the test and background samples, we divided the family counts per GO from each sample, by the samples' average number of GO annotations per family. Statistical significance was numerically assessed by obtaining the expected (adjusted) number of families per GO in 10 000 equally sized random samples derived from the overall population of gene families.

(e) Maximum lifespan and partial correlation coefficients

MLSP recorded for each species was obtained from the animal ageing and longevity database (AnAge) [19]. To correct for the potential contribution of MLSP to the association between GFS and *Ei*, partial correlation coefficients were calculated for each gene family, including MLSP as covariate. The resulting partial coefficient represents the contribution of *Ei* to the variance in GFS which is not explained by variations in MLSP. Only those gene families displaying a significant partial correlation coefficient ($p < 0.01$) between GFS and *Ei* were considered further.

(f) Phylogenetic relatedness test

Phylogenetic generalized least-square approach (PGLS) and maximum-likelihood estimation of λ-values were carried out using the CAPER module in R. Because the parameter λ measures the degree to which the phylogeny predicts the pattern of covariance of a given trait across species (where λ-values close to 0 represent no phylogenetic autocorrelation while values close to 1 represent full phylogenetic autocorrelation) [20–22], this approach allows us to obtain a single accurate measure of phylogenetic autocorrelation for each individual gene family. In order to remove the effect of phylogenetic relationships from our analysis, we determined the parameter λ for each of the 713 gene families with significant partial correlation coefficients for *Ei* and GFS (correcting for MLSP) and eliminated all gene families with a significant phylogenetic interdependence ($p < 0.05$ of $\lambda = 0$, and $p > 0.05$ of $\lambda = 1$). This filtering resulted in 501 gene families on which GO enrichment analyses were subsequently carried out as described above.

(g) Gene expression in human brain

RNA-seq data were obtained for 18 052 genes in a total of 16 human tissues, including brain, derived from the Illumina human body map dataset (Ensembl v. 62). Individual genes were

categorized as *prominently expressed* in the brain if their expression level in this tissue was the highest or second highest among all 16 tissues included (top 12.5th percentile). Over-representation was assessed by counting the number of these genes within a given sample. Statistical significance was assessed by comparing this count with those observed in 10 000 equally sized random samples drawn from the wider pool of gene families.

(h) Co-expression network analysis

Weighted gene co-expression network analysis was carried out based on pairwise Pearson correlations between the expression profiles obtained from the BrainSpan database (http://www.brainspan.org) for over 21 000 genes. Unsupervised hierarchical clustering was used to detect groups, or modules, of highly co-expressed genes following the method described by Zhang & Horvath [23].

3. Results

(a) Gene family size variations in line with encephalization are over-represented in mammals

In order to assess the relationship between encephalization and GFS variations in mammalian taxa, gene family annotations for 39 fully sequenced mammalian genomes were obtained from Ensembl [15]. We included in this study all families with members present in no less than six of the 39 mammalian species (see Methods). This resulted in a total of 12 373 non-overlapping gene families encompassing 595 535 genes, with a mean number of 48.13, and a number of copies per gene family per species ranging from 0 to 448.

Ei for each species was defined as the residual of a log–log least-squares linear regression of brain mass against body mass (see Methods). We obtained correlation coefficients for GFS and *Ei* for each gene family and the resulting distribution of correlation coefficients showed a distinct shift towards positive values (figure 1*a*). A Monte Carlo simulation of the expected distribution based on random permutations of GFS values across species revealed that the observed bias is highly significant ($p \ll 0.0001$). In total, we found 8789 families with $r > 0$, representing a shift of 2602 gene families from the negative to the positive tail of the distribution relative to the expected equal number of positively and negatively correlated families ($\chi^2 = 2189.608$, $p \approx 0$; figure 1*a*, inset). This result demonstrates a highly pronounced over-representation of gene families displaying a positive association between GFS and *Ei*. This observation is not explained by an overall expansion in gene number across species in line with *Ei* ($r = 0.251$, $p = 0.127$), but rather by an over-representation of small gene families among those highly associated with encephalization, combined with few larger gene families displaying decreases in size.

We next asked whether the observed enrichment of *Ei*-related GFS variations was unspecific in terms of the gene populations involved or, alternatively, if this enrichment occurred in gene families specifically associated with certain biological functions. To this end, we used functional GO annotations for 'biological processes' and carried out two complementary tests to assess deviations (from random expectations) in the distribution of GO terms associated with gene families displaying a high correlation between GFS and *Ei*. First, we examined whether there were any significant deviations in the mean correlation coefficient of gene families associated with individual GO terms (see Methods). Out of all 260 functional categories included, only gene families associated with cell–cell signalling, immune response, chemotaxis, neuropeptide signalling pathways and regulation of immune response displayed a significantly higher than expected average correlation values, between GFS and *Ei*, after Bonferroni correction (figure 1*b*). By contrast, no significant bias was observed in functional categories containing families with negative average correlations (not shown).

Second, we measured over-representation of GO terms among the gene families whose GFS variations were most significantly associated with *Ei* ($r > 0$, $p < 0.05$, $n = 1292$). Among these families, we found that GO terms for immune response, chemotaxis, regulation of immune response, female pregnancy, cell–cell signalling, signal transduction, energy reserve metabolic processes, positive regulation of peptidyl-tyrosine phosphorylation and neuropeptide signalling pathways were significantly over-represented after Bonferroni correction (figure 1*c*). No GO terms were found to be significantly over-represented among gene families with the highest negative covariance between GFS and *Ei* (not shown). Taken together, these results show that the observed collective variation in GFS in line with encephalization is not randomly distributed across functional categories but is significantly pronounced in families associated with specific biological functions.

(b) Association between gene family size and encephalization is not explained by lifespan variations

A number of studies on brain evolution have uncovered a robust relationship between relative brain size and lifespan [2,24,25]. In agreement with this, we found a strong association between MLSP and *Ei* among the species included in this study ($r = 0.7912$, $p < 10^{-8}$). Thus, the observed associations between *Ei* and GFS could be secondary to an underlying association between MLSP and GFS. Of the 1292 most significantly correlated families ($r > 0$, $p < 0.05$), 927 displayed a stronger association with *Ei* than with MLSP ($r_{(Ei,\ GFS)} > r_{(MLSP,\ GFS)}$), thereby suggesting a preferential contribution of *Ei* to the observed bias in the correlation distribution ($\chi^2 = 858.74$, $p = 3.3572e^{-187}$, relative to a random equal distribution of stronger associations). GO enrichment analysis was then repeated including only these 927 families revealing a significant over-representation of gene families associated with immune response, chemotaxis, regulation of immune response, energy reserve metabolic processes, female pregnancy, cell–cell signalling, positive regulation of peptidyl-tyrosine phosphorylation and activation of cysteine-type endopeptidase activity involved in apoptotic processes (figure 1*d*). It is worth noting that the complementary GO enrichment analysis carried out on gene families with both the most significant association between MLSP and GFS ($r > 0$, $p < 0.05$) and a stronger association with MLSP than *Ei* ($r_{(MLSP,\ GFS)} > r_{(Ei,\ GFS)}$, $n = 1321$), resulted in no significant enrichment of any GO category. These results shows that enrichment of specific GO terms occurred only among gene families preferentially associated with degree of encephalization, whereas GFS variations potentially associated with increased MLSP showed no significant association with any particular functional category.

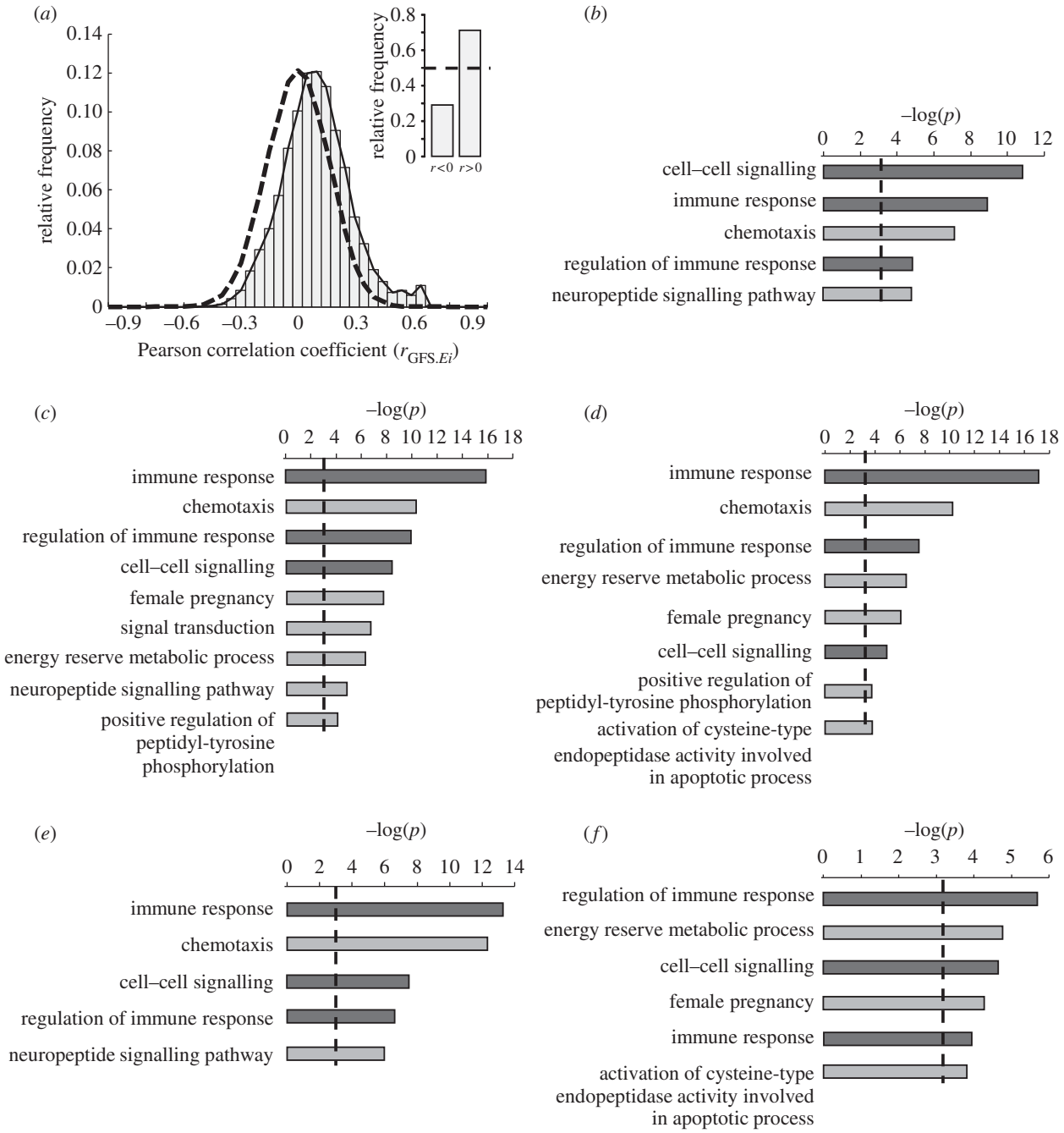


Figure 1. Enrichment of gene family size variations in line with increased encephalization in mammals. (*a*) Histogram showing the distribution of correlation coefficients for GFS and *Ei* in 12 373 gene families encompassing 39 mammalian genomes. A randomization-based estimation of the expected distribution is represented by the dashed line. Inset: distribution of positive and negative correlations relative to the expected distribution (dashed line). (*b*) Deviations from random expectations in the mean correlation coefficient of gene families associated with individual GO terms (expressed as $-\log(p\text{-value})$). Only GO categories with a significant bias are shown. (*c*) Over-representation of GO terms among gene families most significantly associated with encephalization ($p < 0.05$, $n = 1292$). (*d*) GO enrichment analysis among the families displaying the most significant correlation with encephalization after removing all families with a stronger association with MLSP than with *Ei* ($n = 927$). (*e*) GO-terms enrichment analysis among gene families with the most significant positive partial correlation coefficients for *Ei* after controlling for the contribution of MLSP in a multiple regression analysis ($n = 713$). (*f*) GO-terms enrichment analysis among gene families with the most significant positive partial correlation coefficients for *Ei* with no significant phylogenetic interdependence ($n = 501$). Bonferroni-corrected significance thresholds are indicated with a dashed line. Dark bars indicate common GO terms across all five analyses.

Because MLSP may still partly explain the covariance between GFS and *Ei* even if the correlation coefficient of GFS with *Ei* is higher than with MLSP, we used multiple regression analysis to obtain partial correlation coefficients between GFS and *Ei* after controlling for the contribution of MLSP (see Methods). GO terms enrichment analysis was then carried out only among those gene families with the most significant positive partial correlation coefficients (partial $r > 0$, $p < 0.05$, $n = 713$). This analysis revealed a significant enrichment of families functionally associated with regulation of immune response, chemotaxis, cell–cell signalling and neuropeptide signalling pathways (figure 1*e*). These results show that variations in GFS specifically associated with encephalization (i.e. not accounted for by variations in MLSP) are also specifically associated with distinct biological functions.

(c) Phylogenetic relatedness does not explain the observed bias in the distribution of gene families associated with encephalization

For a given gene family, any association between *Ei* and GFS could be the secondary to existing phylogenetic relationships among the species analysed, as in the absence of any selective forces, closely related species will tend to have both similar degrees of *Ei* and similar GFS [20,22]. In order to determine the degree to which phylogenetic effects contribute to the observed shift in the correlation distribution, we used a PGLS approach (see Methods) [20,22]. Out of 713 gene families with the most significant positive partial correlation coefficients between *Ei* and GFS (after correcting for MLSP, see previous analysis), we found a total of 501 gene families for which phylogenetic relationships among species could not account for the covariance between GFS and *Ei*. Among these families, we observed a significant over-representation of gene families associated with regulation of immune response, cell–cell signalling, energy reserve metabolic processes, female pregnancy and activation of endopeptidase activity involved in apoptosis (figure 1*f*). These findings demonstrate that the over-representation of specific biological functions among those gene families most strongly associated with higher *Ei* is neither explained by the known association between MLSP and *Ei* nor by existing phylogenetic relationships among the species analysed.

(d) Gene families with size increases in line with encephalization show expression signatures consistent with brain functions

To assess whether gene family variations in line with encephalization were directly associated with brain function, we characterized the potential relationship between *Ei*-associated GFS variations and patterns of gene expression in the human nervous system. For this analysis, we selected the top 501 *Ei*-associated gene families with both the most significant partial correlation coefficient between *Ei* and GFS and no significant phylogenetic effects (figure 1*f*). Using available expression data from the Illumina human body map (see Methods), we looked at the possible over-representation of genes highly expressed in the human brain within the selected 501 gene families. Individual genes were categorized as *prominently expressed* in the brain if their expression level in this tissue was the highest or second highest among all 16 tissues included (top 12.5th percentile). Statistical significance was assessed by comparing with equally sized random samples drawn from the wider pool of gene families (see Methods). This analysis revealed a significant enrichment, within these gene families, of genes prominently expressed in the brain (figure 2*a*). By contrast, no significant enrichment of genes prominently expressed in the brain was detected among those gene families with the strongest association with MLSP and no significant phylogenetic effects (figure 2*a*).

Genes involved in cortical development have been shown to display higher variance in expression level during the developmental period of the cerebral cortex compared with adulthood [26]. We therefore looked at the possible representation of genes displaying the highest expression variance during human cortical development relative to adulthood, as defined by Sterner *et al*. [26], within the same 501 gene families and found a significant enrichment of genes displaying this pattern of expression (figure 2*b*). By contrast, no significant enrichment of these same genes was observed among the top MLSP-associated gene families (figure 2*b*).

We next asked whether there was any statistical bias in the relative expression of *Ei*-associated gene families across different brain regions. Using human brain RNA-seq data from the BrainSpan dataset (see Methods), we obtained the average expression for each gene in the cortex, subcortical regions or cerebellum and split them into three categories according to the region where the highest average expression was found. This analysis revealed a statistically significant enrichment, among those genes contained within the top 501 *Ei*-correlated gene families, of genes maximally expressed in the cortex (figure 2*c*). No significant enrichment of genes maximally expressed in subcortical regions was observed among these families. By contrast, genes maximally expressed in the cerebellum were found to be significantly under-represented among the top *Ei*-correlated gene families. Taken together, these results reveal that gene families displaying the highest association between *Ei* and GFS are enriched in genes that are prominently expressed in the brain, with maximal expression in the cortex and display an expression signature distinctly associated with cortical development.

In order to characterize further the cortical expression profile of *Ei*-associated gene families, we used a weighted gene co-expression network analysis approach to identify modules of co-expression among genes contained within the top 501 *Ei*-correlated gene families. Using human developmental expression data derived from the BrainSpan dataset, we identified 18 modules (figure 3*a*) associated with distinct temporal patterns of expression. Figure 3*b* shows the time course of expression of six of these modules summarized by the eigengene associated with each module's co-expression matrix. Some of these modules showed the highest expression levels during the early or late fetal period followed by a progressive decline in expression levels with age. This trend may reverse in some instances in late-adult stages (black module, figure 3*b*) or show a progressive increase throughout development as illustrated by the yellow module.

4. Discussion

Our results reveal a highly significant over-representation of gene families displaying a positive association between GFS and level of encephalization. This bias occurs most prominently in families associated with specific biological functions. The most robust and consistent bias was observed in gene families associated with cell signalling, immune regulation and chemotaxis.

While chemotaxis and cell signalling functions are known to play central roles in the nervous system, the significance of the observed enrichment of immune system-associated functions among gene families displaying the highest association between GFS and *Ei* is less clear. In recent years, however, signalling and regulatory mechanisms originally described in the immune system have increasingly been found implicated in key neural-specific roles both in the developing and adult nervous system [27–33]. In addition, in the human cerebral cortex, immune system-related functions have been found to be significantly over-represented among genes displaying higher expression variability in the

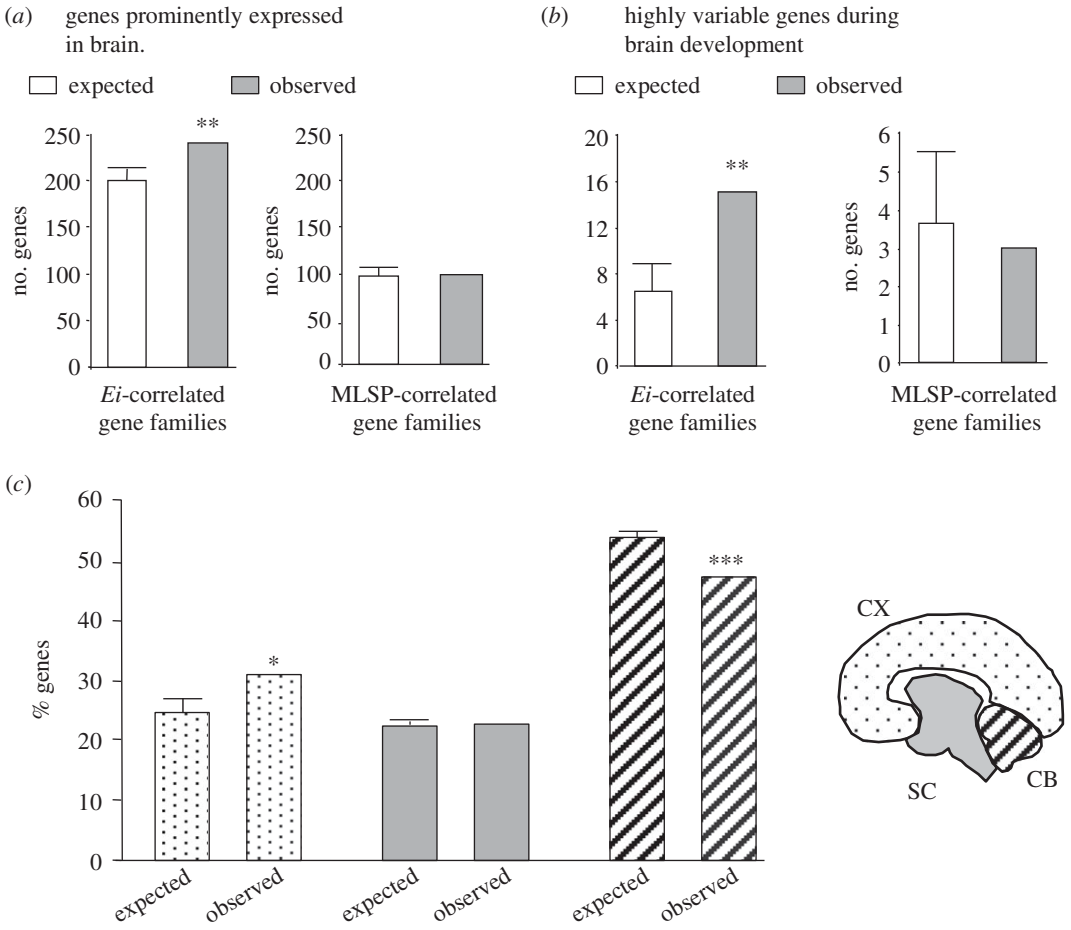


Figure 2. Relationship between *Ei*-associated GFS variations and patterns of gene expression in the human nervous system. (*a*) Over-representation of genes prominently expressed in the human brain (top 12.5th percentile) among the top 501 *Ei* -associated or the top MLSP-associated gene families compared to random expectations. (*b*) Over-representation of genes displaying the highest expression variance during human cortical development relative to adulthood among the top *Ei*-associated or the top MLSP-associated gene families. (*c*) Percentage of genes maximally expressed in the cortical (CX), subcortical (SC) and cerebellar (CB) regions respectively. Expected values (mean $\pm$ s.e.m.) were numerically determined using sized-matched random samples of genes drawn from the wider pool of gene families. *$p < 0.01$; **$p < 0.001$; ***$p < 0.0001$.

developing cerebral cortex than in the adult [26], suggesting a substantial involvement of immune-related signals during cortical development.

Our results, showing a significant over-representation of immune-related functions among *Ei*-associated gene families, support the notion of an underlying and substantial overlap in the regulatory and signalling machinery shared by both the immune and nervous system and in particular during development of the latter.

One possible interpretation is that the observed enrichment of immune-related functions among *Ei*-associated gene families reflects an underlying expansion of immune surveillance in mammals that could be in some way permissive to increased encephalization. While we cannot rule out this possibility, at present, there is little evidence in support of any systematically pronounced and sustained expansion of immune functionalities in mammalian lineages [34]. An alternative interpretation is that signalling and regulatory molecular components that were originally involved in immune-specific functions became gradually recruited by the nervous system in response to the developmental and functional demands of increasingly more complex brains.

The observed association between degree of encephalization and variations of GFS in a large number of gene families is further supported by our finding that *Ei*-associated gene families display a transcriptional signature consistent with brain-specific functions. Indeed, among the gene families most highly correlated with encephalization with no significant phylogenetic effects, we found a statistically significant enrichment of genes prominently expressed in the brain, strongly indicating that these genes are under comparably higher demand in the nervous system relative to other tissues. When restricting the analysis to the relative expression levels within central nervous system regions, we found that these families are enriched in genes prominently expressed in the cortex, suggesting that *Ei*-correlated changes in GFS may have played a substantial role supporting key aspects of cortical evolution. In this regard, it is worth noting that brain evolution in mammalian lineages is characterized by a disproportional expansion of the brain cortex [35,36]. Analysis of the developmental pattern of expression of these families in the human cortex showed that these genes are organized in co-expression clusters or modules with distinct temporal profiles suggesting a substantial involvement of these families in the developmental organization of the brain.

Genes with the highest degree of connectivity within a module are termed hub genes and are expected to be functionally important within the module. By way of illustration, we examined the turquoise module (figure 3*b*) and identified a member of a zinc finger gene family (gene family ID: ENSFM00620000999432) as its main hub gene. Interestingly, all but two of the 20 members of this gene family in humans

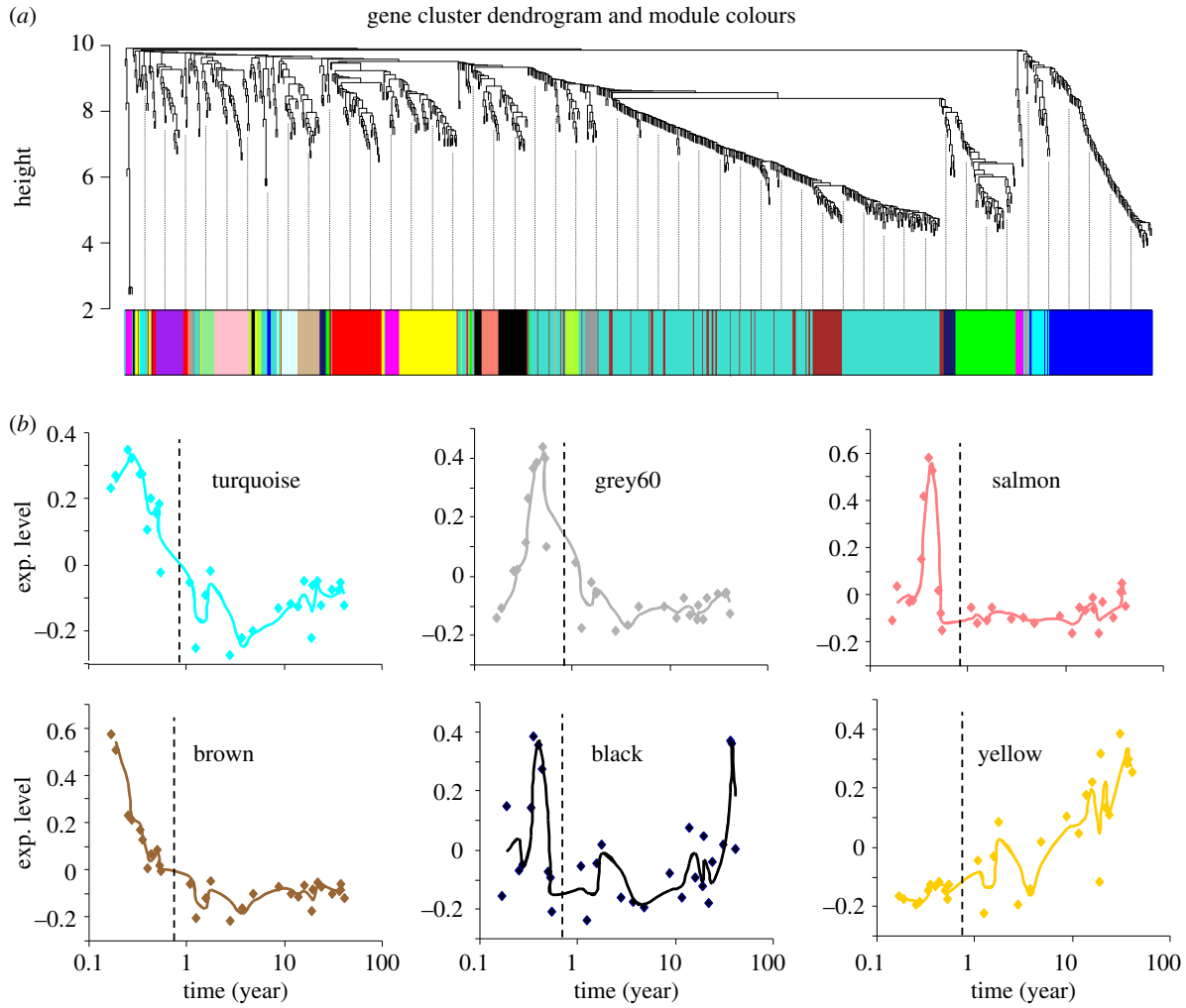


Figure 3. Temporal patterns of cortical expression of *Ei*-associated gene families. (*a*) Weighted gene co-expression network analysis was used to detect co-expression modules among genes contained within the top 501 *Ei*-associated gene families using human brain temporal expression data, revealing 18 co-expression modules (coloured). (*b*) Developmental time course of expression of six representative modules summarized by the level of expression of the eigengene associated with each module's co-expression matrix. Birth point is indicated with a dashed line.

are contained within the same co-expression module. Because genes contained within a co-expression module are thought to be functionally related [23,37], the fact that most members of this zinc finger family are found within the same co-expression module strongly suggests that these genes are functionally related during brain development. We reconstructed the phylogenetic tree of this family and found that the observed pattern is the result of a combination of events of gene loss and gene gain from an original set of four ancestral proteins at the base of the mammalian evolution, overall resulting in a steady increase in the number of gene family members in line with increased level of encephalization ($r = 0.7547$, $p = 2.86 \times 10^{-8}$).

5. Conclusion

In this study, we have found a significant over-representation of GFS variations in line with increased encephalization in mammals. Importantly, this relationship is not accounted for by known correlates of brain size and is not explained by phylogenetic relatedness. The observed bias occurs most prominently in families preferentially expressed in the brain, in particular the cortex, and significantly associated with distinct biological functions.

Based on our results, we propose that variations in GFS associated with encephalization provided an evolutionary support for the specific cellular, physiological and developmental demands associated with increased brain size in mammals.

Acknowledgements. We thank Leticia Ramirez-Lugo from the National University of Mexico for her technical support and helpful comments from two anonymous reviewers.

Funding statement. This work was supported by CONACyT PhD scholarships to A.C.M. and J.M.S., Dorothy Hodgkin Research Fellowship and Royal Society Research Grant to A.U.O. and University of Lincoln support to H.G.

References

1. Roth G, Dicke U. 2005 Evolution of the brain and intelligence. *Trends Cogn. Sci.* **9**, 250–257. (doi:10.1016/j.tics.2005.03.005)
2. Gonzalez-Lagos C, Sol D, Reader SM. 2010 Large-brained mammals live longer. *J. Evol. Biol.* **23**, 1064–1074. (doi:10.1111/j.1420-9101.2010.01976.x)
3. Isler K, van Schaik C. 2006 Costs of encephalization: the energy trade-off hypothesis tested on birds. *J. Hum. Evol.* **51**, 228–243. (doi:10.1016/j.jhevol.2006.03.006)

4. Leonard WR, Robertson ML, Snodgrass JJ, Kuzawa CW. 2003 Metabolic correlates of hominid brain evolution. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **136**, 5–15. (doi:10.1016/S1095-6433(03)00132-6)
5. Weisbecker V, Goswami A. 2010 Brain size, life history, and metabolism at the marsupial/placental dichotomy. *Proc. Natl Acad. Sci. USA* **107**, 16 216–16 221. (doi:10.1073/pnas.0906486107)
6. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, Wyckoff GJ, Malcom CM, Lahn BT. 2004 Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119**, 1027–1040. (doi:10.1016/j.cell.2004.11.040)
7. Shi P, Bakewell MA, Zhang J. 2006 Did brain-specific genes evolve faster in humans than in chimpanzees? *Trends Genet.* **22**, 608–613. (doi:10.1016/j.tig.2006.09.001)
8. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. 2006 The evolution of mammalian gene families. *PLoS ONE* **1**, e85. (doi:10.1371/journal.pone.0000085)
9. Fortna A *et al*. 2004 Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2**, E207. (doi:10.1371/journal.pbio.0020207)
10. Hahn MW, Han MV, Han SG. 2007 Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.* **3**, e197. (doi:10.1371/journal.pgen.0030197)
11. Hughes AL, Friedman R. 2004 Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. *J. Mol. Evol.* **59**, 827–833. (doi:10.1007/s00239-004-0115-7)
12. Rubin GM *et al*. 2000 Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215. (doi:10.1126/science.287.5461.2204)
13. Soshnikova N, Dewaele R, Janvier P, Krumlauf R, Duboule D. 2013 Duplications of hox gene clusters and the emergence of vertebrates. *Dev. Biol.* **378**, 194–199. (doi:10.1016/j.ydbio.2013.03.004)
14. Holland LZ, Short S. 2008 Gene duplication, co-option and recruitment during the origin of the vertebrate brain from the invertebrate chordate brain. *Brain Behav. Evol.* **72**, 91–105. (doi:10.1159/000151470)
15. Ensembl 2012 http://www.ensembl.org. (release 64).
16. Herculano-Houzel S. 2011 Brains matter, bodies maybe not: the case for examining neuron numbers irrespective of body size. *Ann. N.Y. Acad. Sci.* **1225**, 191–199. (doi:10.1111/j.1749-6632.2011.05976.x)
17. Herculano-Houzel S, Collins CE, Wong P, Kaas JH. 2007 Cellular scaling rules for primate brains. *Proc. Natl Acad. Sci. USA* **104**, 3562–3567. (doi:10.1073/pnas.0611396104)
18. Deaner RO, Nunn CL, van Schaik CP. 2000 Comparative tests of primate cognition: different scaling methods produce different results. *Brain Behav. Evol.* **55**, 44–52. (doi:10.1159/000006641)
19. Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhaes JP. 2013 Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D1027–D1033. (doi:10.1093/nar/gks1155)
20. Freckleton RP, Harvey PH, Pagel M. 2002 Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726. (doi:10.1086/343873)
21. Garland Jr T, Bennett AF, Rezende EL. 2005 Phylogenetic approaches in comparative physiology. *J. Exp. Biol.* **208**, 3015–3035. (doi:10.1242/jeb.01745)
22. Pagel M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884. (doi:10.1038/44766)
23. Zhang B, Horvath S. 2005 A general framework for weighted gene co-expression network analysis. *Stat. Appl. in Genet. Mol. Biol.* **4**, 1128. (doi:10.2202/1544-6115.1128)
24. Allen JS, Bruss J, Damasio H. 2005 The aging brain: the cognitive reserve hypothesis and hominid evolution. *Am. J. Hum. Biol.* **17**, 673–689. (doi:10.1002/ajhb.20439)
25. Barrickman NL, Bastian ML, Isler K, van Schaik CP. 2008 Life history costs and benefits of encephalization: a comparative test using data from long-term studies of primates in the wild. *J. Hum. Evol.* **54**, 568–590. (doi:10.1016/j.jhevol.2007.08.012)
26. Sterner KN *et al*. 2012 Dynamic gene expression in the human cerebral cortex distinguishes children from adults. *PLoS ONE* **7**, e37714. (doi:10.1371/journal.pone.0037714)
27. Crampton SJ, Collins LM, Toulouse A, Nolan YM, O'Keeffe GW. 2012 Exposure of foetal neural progenitor cells to IL-1beta impairs their proliferation and alters their differentiation: a role for maternal inflammation? *J. Neurochem.* **120**, 964–973.
28. Gavalda N, Gutierrez H, Davies AM. 2009 Developmental regulation of sensory neurite growth by the tumor necrosis factor superfamily member light. *J. Neurosci.* **29**, 1599–1607. (doi:10.1523/JNEUROSCI.3566-08.2009)
29. Gutierrez H, Hale VA, Dolcet X, Davies A. 2005 NF-kappaB signalling regulates the growth of neural processes in the developing PNS and CNS. *Development* **132**, 1713–1726. (doi:10.1242/dev.01702)
30. Gutierrez H, O'Keeffe GW, Gavalda N, Gallagher D, Davies AM. 2008 Nuclear factor κB signaling either stimulates or inhibits neurite growth depending on the phosphorylation status of p65/RelA. *J. Neurosci.* **28**, 8246–8256. (doi:10.1523/JNEUROSCI.1941-08.2008)
31. McKelvey L, Gutierrez H, Nocentini G, Crampton SJ, Davies AM, Riccardi CR, O'Keeffe GW. 2012 The intracellular portion of GITR enhances NGF-promoted neurite growth through an inverse modulation of Erk and NF-κB signalling. *Biol. Open* **1**, 1016–1023. (doi:10.1242/bio.20121024)
32. Nolan AM, Nolan YM, O'Keeffe GW. 2011 IL-1beta inhibits axonal growth of developing sympathetic neurons. *Mol. Cell Neurosci.* **48**, 142–150. (doi:10.1016/j.mcn.2011.07.003)
33. O'Keeffe GW, Gutierrez H, Pandolfi PP, Riccardi C, Davies AM. 2008 NGF-promoted axon growth and target innervation requires GITRL-GITR signaling. *Nat. Neurosci.* **11**, 135–142. (doi:10.1038/nn2034)
34. Boehm T. 2012 Evolution of vertebrate immunity. *Curr. Biol.* **22**, R722–R732. (doi:10.1016/j.cub.2012.07.003)
35. Nomura T, Gotoh H, Ono K. 2013 Changes in the regulation of cortical neurogenesis contribute to encephalization during amniote brain evolution. *Nat. Commun.* **4**, 2206.
36. Kaas JH. 2013 The evolution of brains from early mammals to humans. *Wiley Interdiscip. Rev. Cogn. Sci.* **4**, 33–45. (doi:10.1002/wcs.1206)
37. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. 2004 Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14**, 1085–1094. (doi:10.1101/gr.1910904)